Towards Measuring and Modeling "Culture" in LLMs: A Survey

Muhammad Farid Adilazuarda[1]*, Sagnik Mukherjee[1]*, Pradhyumna Lavania, Siddhant Singh[2], Alham Fikri A

Abstract

We present a survey of more than 90 recent papers that aim to study cultural representation and inclusion in large language models (LLMs). We observe that none of the studies explicitly define 'culture", which is a complex, multifaceted concept; instead, they probe the models on some specially designed datasets which represent certain aspects of 'culture." We call these aspects the proxies of culture, and organize them across two dimensions of demographic and semantic proxies. We also categorize the probing methods employed. Our analysis indicates that only certain aspects of "culture," such as values and objectives, have been studied, leaving several other interesting and important facets, especially the multitude of semantic domains (Thompson et al., 2020) and aboutness (Hershcovich et al., 2022), unexplored. Two other crucial gaps are the lack of robustness of probing techniques and situated studies on the impact of cultural mis- and under-representation in LLM-based applications. Compilation and details of papers used for the survey can be found via our GitHub repository.

1 Introduction

"Culture is the precipitate of cognition and communication in a human population." - Dan Sperber

Recently, there have been several studies on socio-cultural aspects of LLMs spanning from safety and value alignment (Glaese et al., 2022; Bai et al., 2022b,a) to studying LLMs as personas belonging to certain cultures (Gupta et al., 2024; Kovač et al., 2023) and their skills for resolving dilemmas in the context of value pluralism (Sorensen et al., 2023; Tanmay et al., 2023). https://github.com/faridlazuarda/cultural-llm-papers

In order to make LLMs inclusive and deployable across regions and applications, it is indeed necessary for them to be able to function adequately under different "cultural" contexts. The growing body of work that broadly aims at evaluating LLMs for their multi-cultural awareness and biases underscore an important problem that the existing models are strongly biased towards Western, Anglo-centric or American cultures (Johnson et al., 2022; Cieciuch and Schwartz, 2012; Dwivedi et al., 2023). Such biases are arguably detrimental to the performance of the models in non-Western contexts leading to disparate utility, potential for unfairness across regions. For instance, Haoyue and Cho (2024) and Chaves and Gerosa (2019) show that a conversational system that lacks cultural awareness alienate the users, leading to mistrust and lack of rapport, and eventual abandonment of the system by users from certain cultures. There are also concerns about the impact on global cultural diversity, since if biased models reinforce dominant cultures, whether implicitly or explicitly, they might lead to a cycle of cultural homogeneity (Vaccino-Salvadore, 2023; Schramowski et al., 2021). The recent generation of LLMs, with their impressive ability and widespread availability, only make this issue more pressing. It is therefore a timely moment to review the literature on LLMs and culture.

*Equal contribution

In this work, we survey more than 90 NLP papers that study cultural representation, awareness or bias in LLMs either explicitly (Huang and Yang, 2023; Zhou et al., 2023b; Cao et al., 2024b) or implicitly (Wan et al., 2023). It is quickly apparent that these papers either do not attempt to define culture or use very high-level definitions. For example, a common definition is 'the way of life of a collective group of people, [that] distinguishes them from other groups with other cultures" (Mora, 2013; Shweder et al., 2007; Hershcovich et al., 2022). Not only do the papers typically use broad-brush definitions, most do not engage in a critical discussion on the topic. This is perhaps unsurprising as 'culture" is a concept which evades simple definition.

1.1 Culture in the Social Sciences

Culture is multifaceted, meaning different things to different people at different times. For example, some of the many and often implicitly applied meanings of culture include: (a) 'Cultural Heritage" such as art, music, and food habits (Blake, 2000), (b) 'Interpersonal Interactions" between people from different backgrounds (e.g., ways of speaking in a meeting, politeness norms) (Monaghan et al., 2012), or (c) The "Ways of Life" of a collective group of people distinguishing them from other groups.

There are a variety of sociological descriptions of culture, e.g., Parsons (1972) describes it as the pattern of ideas and principles which abstractly specify how people should behave, but which do so in ways which prove practically effective relative to what people want to do (also see Münch et al. (1992)). However, these too are high-level and hard to concretise. Further complications arise because the instantiation of culture is necessarily situated. Every individual and group lies at the intersection of multiple cultures (defined by their political, professional, religious, regional, class-based and other affiliations) and these are invoked according to the situation, typically in contrast to another group(s).

In anthropology, a distinction has been made between thick and thin descriptions of culture (Geertz, 1973; Bourdieu, 1972). Where culture as understood from the outsiders perspective, e.g. "people of type X believe in Y or behave in a particular manner" is a thin description of culture, as it does not consider the actor's (of type X) personal perception of their context that resulted in that particular belief or the behavior. A thick description of culture, on the other hand, not only documents the observed behaviors but also the actors' own explanations of the context and the behavior, and thus, can capture the insider-view of a culture as captured through people's lived experiences.

Drawing from cultural anthropology, we can frame culture not just as 'the way of life of a people, but as a situated, multi-faceted construct, informed by specific historical and social contexts (Geertz, 1973; Bourdieu, 1972). Employing Geertz's Thick Description approach, future studies should aim to capture not just observable behaviors in different cultural settings but also the lived experiences and internal perspectives that lead to these behaviors. This interdisciplinary engagement with anthropology provides a deeper understanding of cultural nuances, which is critical for LLMs to avoid 'thin' representations of culture.

1.2 Culture in NLP

How then is culture handled in NLP research? As we shall demonstrate, the datasets and studies are typically designed to tease out the differential performance of the models across some set of variables. Before we discuss these, we note that a couple of papers have begun to provide richer definitions of culture. Hershcovich et al. (2022) in their study calls out three axes of interaction between language and culture that NLP research and language technology needs to consider: common ground, aboutness and objectives and values. Aboutness refers to the topics and issues that are prioritized or deemed relevant within different cul-

tures. Common Ground is defined by the shared knowledge and assumptions among people within a culture. Like the sociological and anthropological definitions of culture above, this provides a nice conceptualisation of culture, but practically it is hard to instantiate and measure in NLP studies. A recent survey paper (Liu et al., 2024a) chooses a different definition of culture, based on White (1959) three dimensions of culture: 1) within human, 2) between humans, and 3) outside of human. Based on this, the paper creates a "taxonomy of culture" although the categorisation is a little complex.

In most of the NLP research seeking to examine culture, it is not defined at all beyond the high level. Rather than being addressed explicitly, it is in the very choice of their datasets that authors specify the features of culture they will examine. That is, the datasets themselves can be considered to be proxies for culture. What do we mean by this? The authors of these papers investigating cultural representations in LLMs are seeking to understand how applicable LLMs are to different groups of people and finding them apparently wanting in this count, they then seek to demonstrate and measure this concretely. Whilst they do not define culture beyond the high level (because, we would argue, a practical and actionable single definition of culture is hard to come by), the papers are still measuring some facet or other of cultural differences. The differences that they are measuring are instantiated in their datasets. For example, some papers examine food and drink, others differences in religious practices. These concrete, practical, measurable facets are in effect standing as proxies for culture. Since "cultures" are conceptual rather than concrete categories that are difficult to study directly through computational or quantitative methods, these proxies serve as easy to understand markers of culture that can be concretely captured through NLP datasets.

Given this wholly sensible strategy, it is useful to examine the different instantiations of culture found in this style of research. From food and drink, to norms and values, how have researchers represented culture in and through their datasets? In doing so we make explicit the various facets of culture which have been studied, and highlight gaps in the research. We call for a more explicit acknowledgment of the link between the datasets employed and the facets of culture studied, and hope that the schema described in this paper provides a useful mechanism for this. In addition, we highlight limitations in the robustness of the probing methods used in the studies, which raises doubts about the reliability and generalizability of the findings. Whilst benchmarking is important and necessary, it is not sufficient, as the choices made in creating rigorous benchmarking datasets are unlikely to reveal the full extent of either LLMs cultural limitations or their full cultural representation. Not only is culture multi-faceted, but cultural representation is tied in closely with other related factors such as local language use and local terminology (Wibowo et al., 2023).

Our study also brings out the lack, and the urgent need thereof, for situated studies of LLM-based applications in particular cultural contexts (e.g., restoring ancient texts from ancient cultures (Assael et al., 2022); journalists in Africa (Gondwe, 2023), and digital image making practices (Mim et al., 2024)), which are conspicuously absent from the NLP literature. The combination of rigorous benchmarking and naturalistic studies will present a fuller picture of how culture plays out in LLMs.

The survey is organized as follows. In Section 2, we describe our method for identifying the papers, categorizing them along various axes, and then deriving a taxonomy based on the proxies of cultures and probing methods used in the studies. These taxonomies are presented in Section 3 and Section 4 respectively. In Section 5, we discuss the gaps and recommendations. We conclude in Section 6.

2 Method

Scope of this survey is limited to the study of cultural representations within LLMs and LLM-based applications. Studies on culture in NLP that does not involve LLM have been excluded, and in order to keep this survey focused and manageable, we have also excluded studies on speech and multimodal models.

2.1 Searching Relevant Papers

Our initial step is an exhaustive search within the ACL Anthology database and a manual search on Google Scholar for papers on culture and LLM, with the following keywords: 'culture", 'cultural", 'culturally", 'norms", 'social", 'values", 'socio", 'moral", 'ethics". We also searched for relevant papers from NeuRIPS and the Web Conference. This initial search followed by a manual filtering resulted in 90 papers published between 2020 and 2024. These papers were then manually labeled for (a) the definition of culture subscribed to in the paper, (b) the method used for probing the LLM for cultural awareness/bias, and (c) the languages and the cultures (thus defined) that were studied. It became apparent during the annotation process that none of the papers attempted to explicitly define 'culture." In the absense of definitions of culture, we labelled the papers according to (1) the types of data used to represent cultural differences which can be considered as a proxy for culture (as explained in Sec 1.2), and (2) the aspects of linguistic-culture interaction (Hershcovich et al., 2022) that were studied. Using these labels, we then built taxonomies bottom-up for the object and the method of study.

2.2 Taxonomy: Defining Culture

2.2.1 Proxies of Culture

We identified 12 distinct labels into which the types of data or proxies of cultural difference can be categorized. These can be further classified into two overarching groups:

1) Demographic Proxies: Culture is, almost always, described at the level of a community or group of people, who share certain common demographic attributes. These could be ethnicity (Masai culture), religion (Islamic culture), age (Gen Z culture), socio-economic class (middle class or urban), race, gender, language, region (Indonesian culture) and so on, and their intersections (e.g., Indian middle class).

2) Semantic Proxies: Often cultures are defined in terms of the emotions and values, food and drink, kinship terms, social etiquette, etc. prevalent within a group of people. Thompson et al. (2020) groups these items under "semantic domains", and they describe 21 semantic domains whose linguistic (and cognitive) usage is strongly influenced by culture. We use this framework to organize the semantic proxies of culture.

Note that the semantic and demographic proxies are orthogonal and simultaneously apply to any study. For instance one could choose to study the festivals (a semantic proxy) celebrated in a particular country (a demographic proxy).

2.3 Taxonomy: Probing Methods

There are two broad approaches to studying LLMs - the black-box approach which treats the LLM as a black-box and only relies on the observed responses to various inputs for analysis, and white-box approach where the internal states (such as the attention maps) of the models can be observed e.g. Wichers et al. (2024). Almost all studies we surveyed use the black-box approaches, where typically the input query is appended with a cultural context and presented to the model. The responses of the model are compared under different cultural conditions as well as to baselines where no condition is present. These approaches can be further categorized as:

- **Discriminative Probing**, where the model is expected to choose a specific answer from

a set such as a multiple-choice question-answering setup.

- **Generative Probing** uses an open-ended fill-in-the-blank evaluation method for the LLMs and the text generated by the model under different cultural conditioning are compared.

We have not come across any study on culture that uses white-box approaches, and deem this to be an important gap in the area because these approaches are more interpretable and likely more robust than black-box methods. We present a variety of prompts that are used to probe the model in the black box setting in Appendix A.

3 Findings: Defining Culture

In this section, we discuss how different papers have framed the problem of studying "culture." The findings are organized by the three dimensional taxonomy proposed in Sec 2.2.1 and also presented graphically in Fig 1.

3.1 Demographic Proxies

Most studies use either geographical region (37 out of 90) or language (35 out of 90) or both (17 out of 90) as a proxy for culture. These two proxies are strongly correlated especially when regions are defined as countries (for example, EVS/WVS (2022); Nangia et al. (2020); Koto et al. (2023)). Some of these studies focus on a specific region or language, for example, Indonesia (Koto et al., 2023), France/French (Nangia et al., 2020), Middle-east/Arabic (Naous et al., 2023), and India (Khanuja et al., 2023). A few studies, such as Dwivedi et al. (2023), further groups countries into larger global regions such as Europe, Middle East and Africa. Meanwhile, Wibowo et al. (2023) studied at a more granular province-level Jakarta region, arguing the difficulty in defining general culture even within a country. Typically, the goal here is to create a dataset for a specific region/language and contrast the performance of the models on this dataset to that of a dominant culture (usually Western/American) or language (usually English). This is sociologically problematic, given that there are of course as many different cultural groups and practices in the West as anywhere else. However, for the purposes of these NLP studies, which aim to demonstrate and measure the limited representation of non-Western practices in these models, this approach is practically useful.

Other studies, such as Cao et al. (2023); Tanmay et al. (2023); Quan et al. (2020); Wang et al. (2023) create and contrast datasets in a few different languages (typically 4-8). Very rarely, we see datasets and studies spanning a large number of regions: Jha et al. (2023) proposes a stereotype dataset across 178 countries and EVS/WVS (2022) is a dataset spanning 200 countries; Wu et al. (2023) studies 27 diverse cultures across 6 continents; and Dwivedi et al. (2023) studies social norms of 50+ countries grouped by 5 broad regions. However, almost all studies conclude that the models are more biased and/or have better performance for Western culture/English language than the other ones that were studied.

Of the other demographic proxies, while gender, sexual orientation, race, ethnicity and religion are widely studied dimensions of discrimination in NLP and more broadly, AI systems (Blodgett et al., 2020; Yao et al., 2023), they do not typically focus on cultural aspects of the demographic groups themselves. Rather, the studies tend to focus on how specific groups are targeted or stereotyped by the models reflecting similar real-world discriminatory behaviors. Nonetheless, the persona-driven study of LLMs by Wan et al. (2023) and Dammu et al. (2024) are worth mentioning, where the authors create prompted conversations between personas defined by demographic attributes (cultural conditioning) including gender, race, sexual orientation, class, education, profession, religious belief, political ideology, disability, and region (in the former) and caste in Indian context (in the latter). Analyses of the conversations reveal significant biases and stereotyping which led the authors to warn against

persona-based chatbots in both cases.

In the study of folktales by Wu et al. (2023), where the primary demographic proxy is still region, analysis shows how values and gender roles/biases interact across 27 different region-based cultures. Note that here the object of study is the folktales and not the models that are used to analyze the data at a large scale.

Finally, it is worth mentioning that the range of demographic proxies studied is strongly influenced by and therefore, limited to the "diversity-and-inclusion" discourse in the West, and therefore, misses on many other aspects such as caste, which might be more relevant in other cultural contexts (Sambasivan et al., 2021; Dammu et al., 2024).

3.2 Semantic Proxies

A majority of the studies surveyed (25 papers out of 55 paper on the semantic proxies) focus on a single semantic domain emotions and values from the 21 defined categories in Thompson et al. (2020). Furthermore, there are several datasets and well-defined frameworks, such as the World Value Survey (EVS/WVS, 2022) and Defining Issues Tests (Rest and Kohlberg, 1979), which provides a ready-made platform for defining and conducting cultural studies on values. Yet another reason for the emphasis on value-based studies is arguably the strong and evolving narrative around Responsible AI and AI ethics (Bender et al., 2021; Eliot, 2022).

Of the other semantic domains, Palta and Rudinger (2023) study Food and Beverages where a set of CommonsenseQA-style questions focused on food-related customs is developed for probing cultural biases in commonsense reasoning systems; and Cao et al. (2024b) introduce CulturalRecipes a cross-cultural recipe adaptation dataset in Mandarin Chinese and English, highlighting culinary cultural exchanges.

An et al. (2023) and Quan et al. (2020) focus on named-entities as a semantic proxy for culture, which is not covered in the list of semantic domains discussed in Thompson et al. (2020) but we believe forms an integral aspect of cultural proxy. An et al. (2023) shows that LLMs associate names of people to gender, race and ethnicity, thus implicitly learning a map between names and other demographic attributes. Quan et al. (2020) on the other hand emphasize on the preservation of local named-entities for names of people, places, transport systems and so on, in multilingual datasets, even if these were to be obtained through translation.

Some of the dataset creation exercises have not focused on any particular semantic proxy. Rather, the effort has been towards a holistic representation of a "culture" (usually defined by demographics) through implicitly covering a large number of semantic domains. For instance, Wang et al. (2023) investigates the capability of language models to understand cultural practices through various datasets on language, reasoning, and culture, sourced from local residencies' proposals, government websites, historical textbooks and exams, cultural heritage materials, and academic research. Similarly, Wibowo et al. (2023) presents a language reasoning dataset covering various cultural nuances of Indonesian (and Indonesia).

The absence of culture studies on other semantic domains is concerning, but provides a fertile and fascinating ground for future research. For instance, Sitaram et al. (2023) discusses the problem of learning pronoun usage conventions in Hindi, which are heavily conventionalized and strongly situated in social contexts, and show that ChatGPT learned simplistic representations of these conventions akin to 'thin description" of culture rather than a 'thick", culturally nuanced contextual understanding of the usage. Similarly, the use of quantity, kinship terms, etc. in a language has strong cultural connotations that can be studied at scale.

4 Findings: Probing Methods

The most common approach to investigate cultural representation, awareness and/or bias in LLMs is through black-box probing approaches, where the LLM is probed with input prompts with and without cultural conditions. A typical example of this style is substantiated by the following prompting strategy described in Cao et al. (2023).

> Pick one. Do people in [COUNTRY NAME] believe that claiming government benefits to which you are not entitled is: 1. Never justifiable 2. Something in between 3. Always justifiable

The prompt has two variables, first the [COUNTRY$_{N}$*AME*]*whichprovidestheculturalcontext, andsecond, theinputquestionon'claiminggovernment… truthsfordifferentcultures.Thismethodcanrevealthebiasinthedefaultpredictionsofthemodel, butdoesnotpro… informedwayforcertaincultureifprobedproperly.Mostpaperswesurveyedusesomevariationofthistechnique…*

Furthermore, Kovač et al. (2023) introduces three distinct methods for presenting the cultural context: Simulated conversations, which mimic real-life interactions; Text formats, which involve evaluating responses to various structured text inputs; and Wikipedia paragraphs, where models are tested on their understanding and interpretation of information from Wikipedia articles, offering a diverse set of probing techniques to evaluate model capabilities.

Alternatively, Generative Probing assesses LLMs based on their free-text generation. Evaluating free-text generation is not as streamlined and may require manual inspection. Jha et al. (2023) introduces the SeeGULL stereotype dataset, which leverages the generative capabilities of LLMs to demonstrate how these models frequently reproduce stereotypes that are present in their training data as statistical associations.

Most evaluation techniques use a Single-turn Probing where the cultural context and the probe are given in one go as a single prompt (Tanmay et al., 2023; Ramezani and Xu, 2023). On the other hand, Multi-turn Probing, initially introduced by Cao et al. (2023), evaluates the model's responses over several interactions, allowing for a nuanced understanding of its cultural sensitivity (also see Dammu et al. (2024)).

A limitation of black-box probing approaches is model sensitivity to prompts (Sclar et al., 2023; Beck et al., 2024b) such as the exact wording and format that are irrelevant to the cultural context. This raises questions regarding the reliability and generalizability of the results because one cannot be sure if the observed responses are an artifact of the cultural conditioning or other unrelated factors.

While black-box approaches have been predominant in investigating cultural representation in LLMs, white-box probing methods offer a more interpretable alternative by examining internal model workings to uncover how biases are encoded. Techniques like Gradient-Based Analysis (Wichers et al., 2024; Yu et al., 2023), Attention Mechanism Analysis (Clark et al., 2019), Embedding Space Evaluation (Bolukbasi et al., 2016), and Layer-Wise Analysis (Miaschi et al., 2020) have been primarily applied to bias mitigation-particularly addressing issues like gender and racial biases-within model parameters. However, these studies are currently limited in scope regarding cultural representation; they have not yet been extensively utilized to explore how cultural biases and representations are encoded in LLMs.

5 Discussion and Future Directions

[MISSING/ILLEGIBLE - Content from pages 7-10 was partially missing or restricted in snippet view, however the provided references provide context for the following gaps:]

Lack of a Unified Framework: As noted, the lack of an explicit definition of culture hinders systematic study. Future work should adopt more formal frameworks from sociology and anthropology.

Representational Gaps: There is a heavy focus on Western/Anglo-centric cultures. There is an ur-

gent need for datasets and studies focusing on the Global South and indigenous cultures.

Robustness of Probing: Probing methods are highly sensitive to prompt variations. More robust, perhaps white-box, methods are needed to truly understand a model's cultural knowledge.

Situated Studies: There is a lack of studies on how LLMs impact specific cultural groups in real-world applications.

6 Conclusion

In this paper, we surveyed over 90 papers studying culture in LLMs. We proposed a taxonomy of cultural proxies (demographic and semantic) and probing methods. Our analysis revealed that while values and certain demographics are well-studied, many semantic domains and situated impacts remain unexplored. We hope this survey serves as a roadmap for future research in creating more culturally inclusive and aware LLMs.

References

[1] Adilazuarda, M. F., et al. (2024). Towards Measuring and Modeling 'Culture" in LLMs: A Survey.

[2] Assael, Y., et al. (2022). Restoring ancient texts from ancient cultures.

[3] Bai, Y., et al. (2022a). Training a helpful and harmless assistant with reinforcement learning from human feedback.

[4] Bai, Y., et al. (2022b). Constitutional AI: Harmlessness from AI feedback.

[5] Bender, E. M., et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

[6] Blake, J. (2000). On defining the cultural heritage.

[7] Blodgett, S. L., et al. (2020). Language (technology) is power: A critical survey of 'bias" in NLP.

[8] Bolukbasi, T., et al. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

[9] Bourdieu, P. (1972). Outline of a Theory of Practice.

[10] Cao, Y., et al. (2023). Theory of Mind in LLMs.

[11] Geertz, C. (1973). The Interpretation of Cultures.

[12] Hershcovich, D., et al. (2022). Challenges and Strategies in Cross-Cultural NLP.

[13] Thompson, B., et al. (2020). Cultural variation in enumerative systems.

[14] White, L. (1959). The Evolution of Culture.

[15] Wibowo, H. A., et al. (2023). IndoNLP.

A Sample Prompts

[MISSING/ILLEGIBLE - The full list of prompts mentioned in Section 2.3 was not provided in the source text snippets.]