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

Muhammad Farid Adilazuarda[1*1*2], **Siddhant Singh**, **Alham Fikri Aji**, **Jacki O'Neill**, **Ashutosh Modi**, M

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 underrepresentation in LLM-based applications. Compilation and details of papers used for the survey can be found via our repository.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities across a wide range of natural language tasks. However, these models are known to exhibit biases, often reflecting the data they were trained on. A significant concern is the Western, Educated, Industrialized, Rich, and Democratic (WEIRD) bias, where models perform better for or align more closely with Western cultures.

Understanding how "culture" is represented in LLMs is crucial for ensuring equitable and safe deployment of these technologies globally. In this survey, we analyze over 90 papers dealing with cultural representation in LLMs. We find that the definition of culture in NLP research is often implicit. Researchers typically operate on specific subsets or operational definitions of culture, which we term *proxies*.

Our contributions are as follows:

- We propose a taxonomy of *cultural proxies*, categorized into Demographic Proxies and Semantic Proxies.
- We survey the *probing methods* used to measure these cultural aspects in LLMs.
- We identify significant gaps in current research, specifically regarding the breadth of cultural semantics (e.g., aboutness) and the robustness of measurement techniques.

2 What is "Culture"?

Defining culture is notoriously difficult. In sociology and anthropology, definitions range from shared values and norms to symbols and rituals. Hofstede's cultural dimensions (Indulgence, Long Term Orientation, Masculinity, Uncertainty Avoidance, Power Distance, Individualism) are a common framework. The World Values Survey (WVS) provides another empirical basis for comparing cultures based on shared values.

In the context of NLP and LLMs, we observe that papers rarely adopt a single comprehensive definition. Instead, they focus on measurable attributes.

3 Dimensions of Cultural Proxies

We categorize the operational definitions of culture found in the literature into two main dimensions: Demographic Proxies and Semantic Proxies.

* Equal contribution.

Accepted to EMNLP 2024 Main Conference.

3.1 Demographic Proxies

Demographic proxies define culture based on the identity of the speakers or the region associated with the text.

- **Nationality/Geography:** Using country names or locations as a stand-in for culture. This is the most common proxy. Studies often use datasets labeled by country code (e.g., US, CN, IN).
- **Language/Dialect:** Using the language of the prompt or text to infer cultural context. This relies on the strong correlation between language and culture but can be conflated with linguistic capability.
- **Race/Ethnicity:** Focusing on cultural distinctiveness associated with racial or ethnic groups (e.g., African American Vernacular English).
- **Religion:** Using religious affiliation as a primary cultural marker.
- **Political Leaning:** While distinct from culture, political values often overlap with cultural values in probing studies.

3.2 Semantic Proxies

Semantic proxies focus on the content of the culture—what constitutes the shared knowledge, beliefs, or behaviors.

- **Values:** Abstract ideals that are important to a group (e.g., freedom, tradition). Frameworks like Hofstede or Schwartz are often used here.
- **Norms:** Rules of conduct or social expectations in specific situations (e.g., tipping etiquette, greeting styles).
- **Knowledge/Facts:** Cultural commonsense or specific knowledge about food, festivals, and history (e.g., "Does mapo tofu contain coffee?").
- **Symbols/Rituals:** References to culturally significant symbols or practices.

4 Probing Methods

We categorize the methods used to extract or measure cultural knowledge in LLMs. See Appendix A for detailed black-box probing descriptions.

4.1 Prompting Strategies

- **Direct Inquiry:** Asking the model directly about cultural facts or values (e.g., "What are the values of people in Japan?").
- **Persona/Role-Playing:** Instructing the model to adopt a specific cultural persona (e.g., "Act as a Chinese person...").
- **Contextual Prompting:** Providing context in the prompt that implies a specific culture (e.g., using cultural names or locations).

4.2 Task-Based Probing

- **Cloze Tasks / Masked Language Modeling:** Predicting missing words that carry cultural significance.
- **Multiple Choice QA:** Answering questions from surveys like WVS or specific cultural datasets (e.g., CANDLE).
- **NLI / Sentiment Analysis:** Assessing if the model's judgments on entailment or sentiment align with cultural annotators.

5 Results and Analysis

[Note: Summarizing general findings from the surveyed papers]

Western Bias: Most studies confirm that off-the-shelf LLMs (especially those trained primarily on English data) exhibit a strong alignment with Western, specifically American, values.

Effect of Language: Prompting in a specific language often shifts the model's outputs towards the culture associated with that language, known as the "multilingual-cultural correlation." However, this is not consistent across all languages or models.

Stereotyping: While persona prompting can improve cultural alignment, it also increases the risk of the model generating caricatures or offensive stereotypes.

6 Discussion and Limitations

6.1 The "Aboutness" Gap

Current research heavily focuses on values and norms (subjective culture). There is a lack of focus on "aboutness"—the topics, entities, and domains

that are salient in a specific culture (Hershcovich et al., 2022). For example, knowing that "cricket" is a culturally central topic in India beyond just the rules of the game.

6.2 Robustness of Probing

Probing results are highly sensitive to prompt phrasing. A slight change in the question can lead to significantly different cultural alignments. This lack of robustness makes it difficult to draw definitive conclusions about "what the model knows."

6.3 Situated Studies

There is a scarcity of studies investigating the downstream impact of cultural misalignment in real-world applications. Most work remains at the level of intrinsic evaluation (probing).

7 Conclusion

We surveyed over 90 papers on cultural representation in LLMs. While progress has been made in identifying biases and measuring alignment with values (e.g., WVS), the field lacks a unified definition of culture. Future work should expand beyond simple demographic proxies, explore semantic domains like "aboutness," and improve the robustness of measurement techniques.

Acknowledgements

[ILLEGIBLE / MISSING in snippets]

A Black Box Probing Methods

[Image 10 in source]

Black box probing refers to methods that do not require access to the model's internal weights or gradients. These are essential for evaluating proprietary models like GPT-4 or Claude. Common techniques include:

1. **Survey Filling:** Feeding questions from social science surveys (e.g., WVS, Pew) to the LLM and comparing the distribution of answers to human responses.
2. **Vignette Analysis:** Presenting short stories or social situations and asking the model to judge the appropriateness of actions.
3. **Cultural NLI:** Using Natural Language Inference datasets where premises and hypotheses are culturally grounded.