DOSA: A Dataset of Social Artifacts from Different Indian Geographical Subcultures

Agrima Seth[1], **Sanchit Ahuja**[2], **Kalika Bali**[2], **Sunayana Sitaram**[2]
[1]School of Information, University of Michigan,
[2]Microsoft Research India
agrima@umich.edu, {t-sahuja, kalikab, sunayana.sitaram}@microsoft.com

Abstract

Generative models are increasingly being used in various applications, such as text generation, commonsense reasoning, and question-answering. To be effective globally, these models must be aware of and account for local socio-cultural contexts, making it necessary to have benchmarks to evaluate the models for their cultural familiarity. Since the training data for LLMs is web-based and the Web is limited in its representation of information, it does not capture knowledge present within communities that are not on the Web. Thus, these models exacerbate the inequities, semantic misalignment, and stereotypes from the Web. There has been a growing call for community-centered participatory research methods in NLP. In this work, we respond to this call by using participatory research methods to introduce *DOSA*, the first community-generated **D**ataset **o**f 615 **S**ocial **A**rtifacts, by engaging with 260 participants from 19 different Indian geographic subcultures. We use a gamified framework that relies on collective sensemaking to collect the names and descriptions of these artifacts such that the descriptions semantically align with the shared sensibilities of the individuals from those cultures. Next, we benchmark four popular LLMs and find that they show significant variation across regional sub-cultures in their ability to infer the artifacts.

Keywords: Generative AI, LLMs, social artifacts, human-centered dataset creation, participatory research, Global South, non-western dataset

1. Introduction

Large Language Models (LLMs) are increasingly being integrated with various applications that have a direct social impact (Tamkin et al., 2021), such as chatbots for health advice (Jo et al., 2023; Cabrera et al., 2023) and content moderation (Wang et al., 2023). There has been an increase in the concerns about what cultural nuances they have, what world knowledge they encode, what ideologies their outputs mimic (Atari et al., 2023), and what gender (Thakur, 2023; Kotek et al., 2023), race (Fang et al., 2023), political identities (Motoki et al., 2023), and experiences are these models aware of. LLMs trained on large-scale, diverse, and filtered web data are considered by many as adept in performing multiple tasks (Brown et al., 2020). However, the Web itself is lacking and inequitable, i.e., while certain cultures and their related knowledge are represented more than others, the knowledge of many cultures is missing altogether. Various scholarships on a wide variety of tasks, such as question-answering (Palta and Rudinger, 2023), value alignment, and fairness, have shown that these models predominantly align with the Western, Educated, Industrialized, Rich, and Democratic (WEIRD) ideologies (Atari et al., 2023), are Anglo-centric and reproduce some harmful stereotypes and biases (Thakur, 2023; Kotek et al., 2023; Abid et al., 2021)

For LLMs to have global acceptability, we must understand what cultural knowledge and behaviors

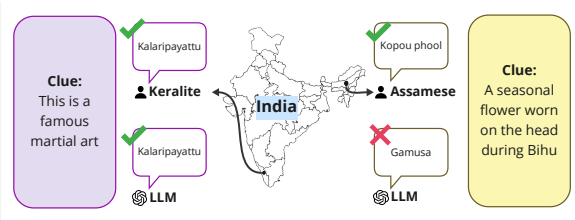


Figure 1: Are LLMs equally aware of social artifacts from different subcultures within a country? For social artifacts from different state-based subcultures of India, we prompt LLMs with unique information that differentiates a social artifact from others and evaluate their overall accuracy for each state.

their outputs mimic. There are multiple dimensions to understanding the cultural acceptability of LLMs. One dimension that has been studied in the past to assess LLMs applicability for decision-making is cultural values like individualism and collectivism, authority and subversion (Hofstede, 2011; Graham et al., 2013; Seth et al., 2023). However, (a) for tasks on content production like text generation and creative writing, (b) to avoid risks like cultural erasure by omission and propagating hegemonic views language models, and (c) to not place *extra burden* of communication on members of non-Eurocentric cultures, these models need to be aware of the social artifacts and the commonplace knowledge associated with them that is present in the target society and actively use this knowledge in

content production (Prabhakaran et al., 2022) [1]. However, this dimension of cultural awareness and alignment is understudied.

Surprisingly, there has not been a systematic evaluation of whether and how the existing models encode the knowledge about artifacts considered essential and commonplace by individuals from the culture. There are many challenges in creating robust evaluation datasets for cultural understanding. First, getting data across different cultures is difficult from the Web because not all cultures find equal representation on the Web. Even when artifacts are represented on the Web, they are likely to be those that have been embraced or recognized by other mainstream societies and many-a-times are remnants or reproduce colonial knowledge (Mamadouh, 2020), which further diminishes the voices of the community members and might propagate stereotypes (Qadri et al., 2023). The second challenge is accessing knowledge sources and people from whom cultural data can be collected and determining how to meaningfully involve individuals from different cultures in dataset creation and multicultural research. While past work in multilingual studies created parallel datasets using translation as a strategy, some culture-specific concepts and objects often do not have equivalents in other cultures and hence either do not have a linguistic equivalent or the semantics do not correlate with the sensibilities of community members(Hershcovich et al., 2022). Thus, the creation of culture-based datasets and subsequent evaluations of LLMs is a challenging task.

In this work, we use bottom-up, community-centered participatory research methods in a non-Western context and engage with community members to introduce a dataset of 615 social artifacts' names and descriptions across 19 regional subcultures of India. We use surveys and implement a gamified framework to create a dataset of social artifacts and use this dataset to benchmark the cultural familiarity of the four most widely used and recent LLMs - GPT4(OpenAI, 2023), LlAMA2(Touvron et al., 2023), PALM 2(Anil et al., 2023), and FALCON (Almazrouei et al., 2023). In particular, this work focuses on the diverse culture of India - a country in the Global South.

Contributions: Past work positions that language technologies can benefit from integrating community intelligence via participatory research (Diddee et al., 2022). However, ***"how"*** one designs participatory frameworks for effective involvement of users is a non-trivial problem. In this paper, we respond to the calls for more community-centered research and show how participatory research methods can effectively create datasets. Second, this is the first paper that explores cultures at a geographical level within a country - India and presents a social artifact dataset to expand the field's understanding of the cultural familiarity of LLMs. Next, we benchmark four widely used LLMs (both open and closed sources) and find a significant inter-LLM variation in their familiarity with the social artifacts. Third, we discuss the obstacles and learnings derived from engaging with participatory research to create evaluation datasets. Thus, our work offers an example of how technology evaluation can benefit from engaging community members using participatory research.

Culture is a complex societal-level concept, and it can be defined by multiple factors: location, sexuality, race, nationality, language, religious beliefs, ethnicity, etc. Past work has shown the significance of geographic boundaries in determining cultural identity, such as the World Value Survey (Inglehart et al., 2018). This study focuses on studying the cultural identities based on India's geographic states. India has 28 geographic states, each with different languages, food, and customs, many of which also do not find appropriate representation on the Web. This is the first attempt to use participatory research to collect social artifacts that are commonplace and perceived as important by the members of the respective communities. While this is not a complete dataset of all important social artifacts, future work could draw from this paper to design participatory research-based methods to scale the dataset to more subcultures.

2. Related Work

2.1. Cultural Awareness of Language Models

Past work on LLM evaluations has focused on understanding the personality, values, and ethics encoded in these models. These works have probed these language models using established psychometric and cultural instruments like IPIP-NEO and the Revised NEO Personality Inventory (Safdari et al., 2023), Moral Direction framework, Moral Foundation Theory, Hofstede's cultural dimensions, and Schwartz's cultural value (Yao et al., 2023; Schramowski et al., 2022; Arora et al., 2022; Hämmerl et al., 2022; Fischer et al., 2023). Some past works have also leveraged questions on ethical

[1] For example, "Dosa," a crispy, savory dish in southern states of India, might be called a crepe-like dish; however, the two are not equivalent, and the community where Dosa comes from would not use the linguistic formation "crepe-like" to describe it. Yet the Wikipedia definition takes a Euro-centric stance to define this social artifact https://en.wikipedia.org/wiki/Dosa_(food)

dilemmas (Tanmay et al., 2023) to elicit what human values LLMs mimic in their output. Prior work in fairness has looked at the stereotypes and biases encoded in these models towards specific communities (Thakur, 2023; Kotek et al., 2023; Abid et al., 2021). In contrast to these works, which look at what ideologies LLMs have learned and the biases in the training data for LLMs, our work focuses on creating a community-centered dataset of artifacts across different regional sub-cultures to evaluate LLMs' knowledge and its alignment with the shared body of knowledge that is held in common and considered important by the members of the respective cultural communities.

The most closely related work to ours is Acharya et al. (2020), which used surveys to get information on four specific rituals from MTurkers in the USA and India, and Nguyen et al. (2023), which scrapped Wikipedia to create a knowledge graph of commonsense knowledge of geography, religion, occupation and integrated it with LLMs. However, to the best of our knowledge, ours is the first work that leverages bottom-up participatory research to build a dataset of social artifacts that evaluates the alignment of commonplace knowledge of community members and LLMs and does not use pre-defined categories to restrict its participants.

2.2. Participatory Research for Dataset Creation

Participatory research argues that individuals who are affected by technology should be involved in designing and evaluating it. Human-computer interaction (HCI) as a field has extensively used surveys, focus groups, and interviews, but surprisingly, the use of participatory methods in NLP is largely lacking (Diddee et al., 2022). Prior work in knowledge elicitation has shown success in using "games with a purpose (GWAP)" for collecting commonplace knowledge and verifying concepts and their relations (Balayn et al., 2022; Von Ahn et al., 2006). An essential aspect of social artifacts is the shared knowledge and understanding of the artifact's significant aspects, use, and unique and differentiating characteristics (Refer Fig 1) (Stephenson, 2023). Furthermore, leveraging GWAP allows us to collect more implicit knowledge based on concepts and mental models that would be otherwise hard to codify in a formal written language, usually found in datasets based on scrapping data from the Web. Thus, in this work, we use two methods of participatory research - Surveys and GWAP. Drawing inspiration from prior GWAP, we formulate a modified version of the classic Taboo game (Refer Section 3.4.2) to elicit cultural artifacts and knowledge. While the previously proposed games (Balayn et al., 2022) restrict the players to specific templates, in our work, we relax this requirement by allowing them to formulate clues in natural speech.

3. Methodology for the Dataset Creation

To record information grounded in the community's shared knowledge that differs from the typical article-like written knowledge found in datasets crawled from the Web, we collect the data of social artifacts by combining two participatory research methods - *Survey and Games with a Purpose (GWAP)*. First, we administer the survey to participants across the 19 Indian States, asking them to self-identify their cultural identity and name social artifacts considered important in that subculture. Next, we use these artifacts to design a GWAP and recruit participants from these 19 Indian States to provide information about the artifacts. The game also primes the participants, and when asked to volunteer to share more artifacts and their descriptions, post-game helped us expand the dataset to 615 artifacts across 19 states.

3.1. Target Cultures

India has 28 states and 8 Union Territories, each with diverse food, handicrafts, dance forms, festivals, rituals, and practices. There is a myriad of operationalizations for the construct of 'culture,' and combining each is beyond the scope of this work. This study focuses on studying the cultural identities based on India's geographic states because the states in India were demarcated on the lines of linguistic and ethnic identities (States Reorganisation Act, 1956)[2], which have a strong correlation to a shared sense of culture in individuals from similar geographic regions (Singh and Sharma, 2009). Our study relied on the population data from the World Value Survey (WVS) Wave 6 (Inglehart et al., 2018) in 2014. The survey was conducted in 18 States of India, stating that 95% of India's population resided in those 18 states. However, the state of Telangana is a new state that was formed after the last WVS was conducted in India. Hence, we decided to collect data from 19 states — the 18 states mentioned in the WVS and the state of Telangana.

3.2. Pilot Study

For the pilot experiment, we recruited 14 participants. It was an in-house pilot with employees at our research lab as the participants. Participants were asked to self-identify the state that best represents their cultural identity. In the pilot study, we

[2]https://pwonlyias.com/upsc-notes/state-reorganisation-act/

had participants from 6 geographic cultures. First, these participants took the survey questionnaire and then engaged with the gamified framework. The participants were compensated for their time. Using the learnings from our pilot, we iterated over our survey questions and game design and finalized our methodology to gather data at a larger scale.

3.3. Survey Questionnaire

We used Karya Inc [3], an ethical data company that engages economically disadvantaged Indians in digital work, to asynchronously administer five surveys across the 19 states in India, making it a total of 95 surveys. Due to operational issues, data from Madhya Pradesh could not be gathered. Each state has its official language. For example, Kannada is the official language of Karnataka, while Hindi and Urdu are the official languages of Uttar Pradesh. Past work has shown that users in India increasingly use the Roman script for online communication (Ahmed et al., 2011; Gupta et al., 2012). Hence, we administered the survey in English and restricted the survey to participants with at least 12 grades of education and a working or advanced level of fluency in English. The answers for social artifacts were Romanized, i.e., written in Latin script. The compensation was decided according to the cost of living in India, and each participant was compensated with an Amazon gift voucher of INR 500 [4](See Appendix A.1 for the complete survey).

As discussed earlier, cultural identity is a complex concept. To increase the clarity of instructions, provide more context to the question, and give examples of social artifacts, we added audio instructions providing more details on what cultural identity and social artifacts meant. In the survey, the survey takers were asked to self-identify the state in India that best represents their cultural identity and three states that they believe are culturally similar to theirs. Next, the participants were asked to list five social artifacts that they believe are important to their cultural identity and would be known to a reasonable number of people who share a similar geographic cultural identity. We rejected surveys where the survey takers marked the same states as the most and least similar, provided the same answers with other participants verbatim, or reiterated the same artifacts given in the instructions as examples. To help make up for the rejected surveys, the survey was re-administered. Overall, we collected 267 artifacts across all 18 states.

[3] https://www.karya.in/

[4] https://en.wikipedia.org/wiki/List_of_countries_by_minimum_wage

3.3.1. Data Cleaning and Processing

The survey design lends most artifacts to be transliterated, and the open-ended questions on social artifacts allow the same artifacts to be listed in varying ways. Thus, we manually reviewed the responses. At the end of this phase, the artifacts collected from the survey cover categories like names of local food cuisines, landmarks, rituals, textiles and handicrafts, dance and music forms, and literary or important political figures (Refer Table 6 for examples of the artifacts). The details are in Appendix A.2.

3.4. Knowledge Extraction

As members of a shared cultural community, we develop shared concepts and understandings and use them when communicating with other members of the same community. These shared concepts and understandings manifest in the language we use for communicating about these artifacts, which may vary from the more formal written information found in traditional data sources. For example, instead of referencing a famous landmark (Kashi Vishwanath Temple), we might refer to a well-known movie (Don) to communicate the place (Benaras) we are talking about. Through this game, we aim to collect knowledge about these artifacts such that it is grounded in the shared concepts and understandings of the community.

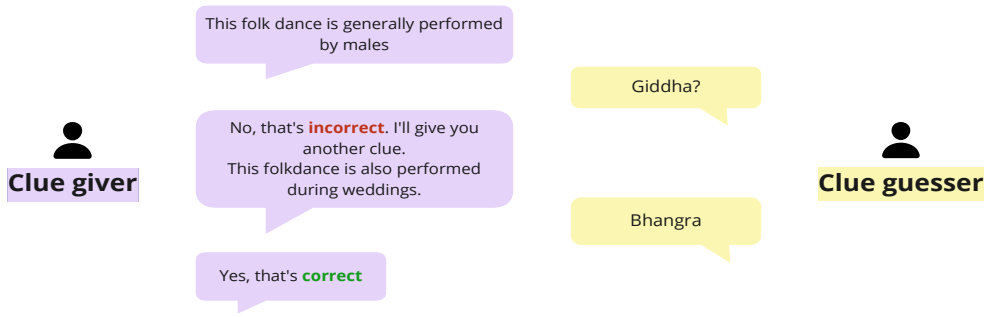
3.4.1. Recruitment and Onboarding for the Game

We recruited a new set of participants for this phase by broadcasting email calls to 2 educational institutes in India and reaching out to friends of friends (including parents). These participants were also asked to self-identify the state in India that best represents their cultural identity. Based on their responses, the participants were paired with another participant from the same culture. We recruited six participants from each of the 18 states to participate in the game. Each participant was compensated with an Amazon gift voucher of INR 500, and the game duration ranged between 1 hour - 90 mins. The list of artifacts obtained from the Survey (Section 3.3) for the corresponding state was used to conduct the game.

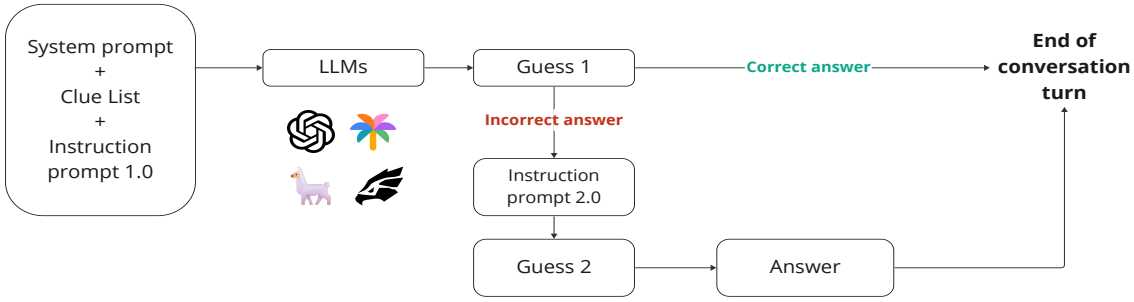
3.4.2. Game Mechanics

Initialization: At the start of the game, we shuffle the artifacts obtained for that state from the survey. Then, both participants are given a mutually exclusive list of a near-equal number of artifacts.

Playing the game: Each player had two roles: (a) the one who gives the clues for the artifact - the



(a) Illustration of the game mechanics and the knowledge elicited from the participants (from Punjab).



(b) Illustration of how LLMs were prompted for benchmarking.

Figure 2: Illustrations highlighting the game mechanics for knowledge elicitation and LLM benchmarking for cultural familiarity

CLUE GIVER and (b) the one who tries to identify what artifact is being talked about - the GUESSER. In each turn, the players alternate their roles. For example, let's call the players A and B. If, in the first turn, player A is the CLUE GIVER, then player B is the GUESSER, and in the next turn, player B becomes the CLUE GIVER, and player A becomes the GUESSER (Figure 2a). The main goal for each player is to use the clues the opponent gave to guess the name of their artifact. To help elicit the most important, unique, and differentiating information about the artifact, the rules of the game were:

1. The CLUE GIVER could give a maximum of 5 clues in the form of a simple sentence.
2. The CLUE GIVER could not use words synonymous with the artifact.
3. The GUESSER had only two chances to make a guess.
4. The GUESSER could not ask any clarification questions.

During the game, the researchers primarily served as observers and ensured enforcement of the above-stated rules.

Clue formulation: While we did not restrict the clues to be given in a templatized form, we did ask the CLUE GIVER to highlight the information that (a) most people with the shared geographic culture would be aware of or agree with, and (b) can be considered the most defining and distinctive to it. Not being restricted by a templatized format for clue formation allowed the players to generate a very rich dataset of both positive or generative knowledge (i.e., what the artifact is) and negative or discriminative knowledge (i.e., what the artifact is not).

3.5. Post-processing of the Artifact Descriptions

Since we recruited six participants from each culture, each artifact was described by three participants. After each game, we transcribed the clues given by the CLUE GIVER.

Across the games for a State, we observed that the clues given to identify the artifact were majorly the same, with minor differences in the wordings. After the CLUE GIVER finished giving clues and the GUESSER made their final guess, we also asked the GUESSER to rate the accuracy and quality of the clues. The perceived quality from the GUESSER and the saturation of the information in the clues over the multiple games allow us to claim reasonable validity and comprehensiveness about the artifacts' descriptions.

3.6. Expanding the Artifact Dataset

To expand the list of social artifacts after the game, we asked the participants, "What other social artifacts can be added to the list? And what would be the best way to describe them if they were a part of the list used in the game?" Since this list was expanded by conversing simultaneously with two individuals from the same culture, there was always an implicit quality check because another participant verified the artifact's name and description. Over the multiple rounds of the game, we observed that the artifacts mentioned by the participants were majorly the same. We refer to these new artifacts and their descriptions as "expanded dataset". We will release both the original and expanded dataset of artifacts for the research community to use.

3.7. Dataset Characteristics

Participant diversity: This study was conducted by actively recruiting individuals from 18 different Indian states. To maximize the diversity within the regions, we used Karya's platform to distribute the survey, allowing us to reach participants from lower socio-economic backgrounds in urban and rural settings. To maximize the diversity of the participants for the game, we recruited participants who are (a) attending public universities and (b) local volunteers from various NGOs. Indian public universities have a mandate for affirmative action under which they provide "reservations" based on social and caste categories. This ensures a high diversity of participants based on gender, socio-economic status, and caste. Similarly, engaging with NGO volunteers, especially in States with a significant population of Tribes, ensures the diversity of participants.

Diversity of Artifacts across different Identities: Since we used a free-form text-based survey to collect the names of artifacts and expanded them using a semi-structured interview, the artifacts collected followed a bottom-up approach, unlike prior works that relied on a more top-down approach. Hence, the DOSA dataset generated from community-centered participatory research was not limited to some pre-defined categories but encompassed a broad range of categories. Next, the diversity of participants in the survey and the game helps us ensure that the artifacts are representative of more than one community within the geographic regions.

States	Original Artifacts	Expanded Artifacts	Total
Andhra Pradesh	14	13	27
Assam	11	56	67
Bihar	12	6	18
Chhattisgarh	9	10	19
Delhi	10	0	10
Gujarat	18	15	33
Haryana	12	17	29
Jharkhand	21	12	33
Karnataka	19	16	35
Kerala	16	13	29
Maharashtra	16	18	34
Odisha	12	32	44
Punjab	20	25	45
Rajasthan	11	6	17
Tamil Nadu	20	29	49
Telangana	13	28	41
Uttar Pradesh	13	34	47
West Bengal	20	18	38
Total			**615**

Table 1: Statistics regarding the number of original (from survey) and expanded (post-in-person game) artifacts available in DOSA.

4. Benchmarking LLMs Cultural Familiarity

We investigate whether the underlying data that large language models were trained on gives them enough context to be familiar with the social artifacts in the DOSA dataset. To make it comparable to the game (Section 3.4) that humans played, we prompt the models with descriptions of the artifacts from Section 3.4 and measure their GUESSING accuracy. In this section, we summarize our experimental setup, models, and the evaluation strategy used in this work to benchmark four LLMs for their cultural familiarity.

4.1. Experimental Setup

We chose a balanced mix of "popular" open source (Llama-2 (Touvron et al., 2023), Falcon (Almazrouei et al., 2023)) and closed source (GPT-4 (OpenAI, 2023) and Palm 2 (Anil et al., 2023)) models to benchmark their cultural familiarity. We chose the best variants of these models that could be supported by the available compute resources, i.e., a single A100 GPU machine. We use the chat and instruct variant of Llama-2 with 13 billion parameters (Touvron et al., 2023) and Falcon with 7 billion parameters (Almazrouei et al., 2023) as the open-source models. We build an interface using Langchain (Chase, 2022) to interact with these models, and to ensure reproducibility, we keep the model temperature at 0. Llama-2 and Falcon were loaded on an A100 GPU and inferred using 16-bit quantization. We used langchain to simulate a chatbot wherein we give our models the initial **System Prompt** with the clue list and the **Instruction Prompt 1.0**. If the FIRST GUESS is correct, the conversation turn ends, and we move on to the

next artifact. If the answer is incorrect, **Instruction Prompt 2.0** mentioning the incorrectness is given, and we ask for a SECOND GUESS (Figure 2b). The prompt templates are constant across models except for the special tags (from (Touvron et al., 2023) and the model documentation (Almazrouei et al., 2023)) appended to the Llama 2 and Falcon models for them to work as intended. The prompts were designed to ensure they are as close to the game conducted with the human participants to collect information about the social artifacts and ensure that the LLM has enough context to predict them.

System Prompt: You are an agent who is well-versed in the cultures of the world. You are playing a game of taboo with another agent who is also well-versed with the cultures of the world. You can only make two guesses to identify this social artifact correctly, and you cannot ask any clarification questions. Social artifacts are objects that help us connect and stay associated with the culture. These objects are known and have significance to most people who consider themselves as a part of that culture and serve as a way of identifying themselves with the culture and the people in that culture. Your clues are: {CLUELIST}

Instruction Prompt 1.0: Name the object based on the above clues from {STATE}. I do not need to know your reasoning behind the answer. Just tell me the answer and nothing else. If you do not know the answer, say that you do not know the answer. Format your answer in the form of ANSWER: your_answer_here.

Instruction Prompt 2.0: Your first guess is not correct. While making your second guess, please stick to the format as ANSWER: your_answer_here.

Figure 3: The instructions used for prompting the LLMs.

4.2. Evaluation Setup

We use accuracy as the primary evaluation metric for assessing the LLMs' cultural familiarity. We report accuracy at three levels - accuracy@GUESS1, accuracy@GUESS2, and overall accuracy. Since LLMs may sometimes produce transliterations that differ from the actual ground truth to ensure the validity of measurement to ascertain the correctness of the guess, we manually matched both the first and second guesses against the ground truth.

1. accuracy@GUESS1: To assess how often the model accurately predicts the artifact on the first attempt, it is calculated by dividing the number of correct FIRST GUESS predictions by the total number of social artifacts.

2. accuracy@GUESS2: This is calculated by dividing the number of correct SECOND GUESS predictions by the total number of unguessed artifacts (i.e., artifacts not predicted correctly at GUESS1).

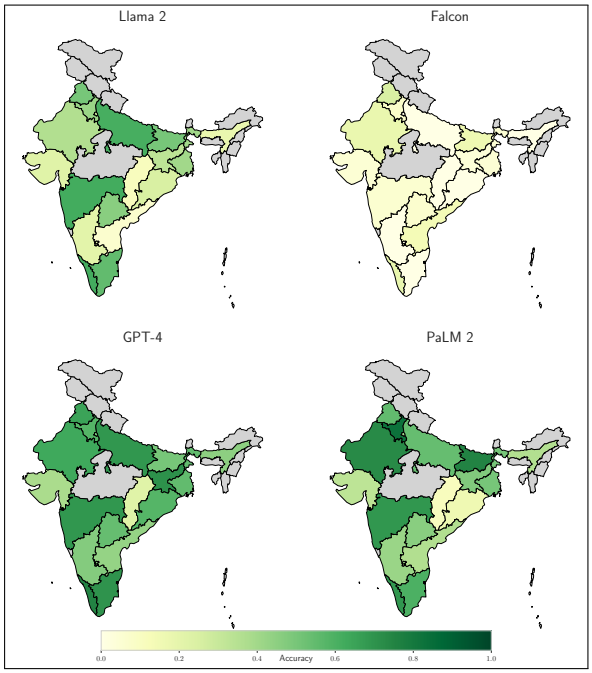
3. Overall Accuracy: To quantify how often the model correctly predicts social artifacts (both the first and second guess, if applicable). It is calculated by dividing the number of correct predictions by the number of social artifacts.

5. Results

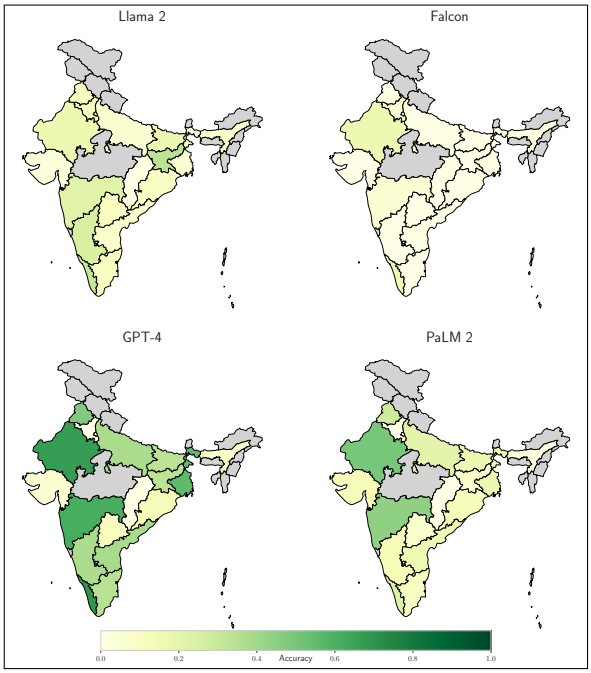
First, the LLMs were evaluated for cultural familiarity with the artifacts collected from the survey (Section 3.3). Next, we evaluated their cultural familiarity with the artifacts from the expanded dataset (Refer Table 4). We find that apart from the well-studied Anglo-centrism of LLMs, they significantly varied in their familiarity with regional subcultures in India, as seen in Figure 4. GPT-4 and Palm 2 perform significantly better than their open-source counterparts - Llama 2 and Falcon, with Falcon barely making any correct guesses. While GPT-4 has performed better overall, Palm 2 performs better for Bihar, Haryana and Rajasthan. We also find that across all cases, the SECOND GUESS does not lead to an increase in accuracy, implying that even when the feedback is given to the models that they are wrong, they are unable to correct themselves and get to the right artifact - an effect more pronounced for the open-source model Falcon.(Refer table 3 for accuracy@GUESS1 and accuracy@GUESS2) Further, we also see that these models do not perform equally on data from each state; for example, although GPT-4 is the best-performing of all the four, it does not work at equal accuracy across all states Fig 4a. For instance, it still cannot identify half of the artifacts from states that rank higher on the Multidimensional Poverty Index, like Assam and Chhattisgarh (Aayog, 2023). We then evaluated the LLMs on the expanded artifacts and found a sharp decrease in the models' performance (Refer Fig 4b). One hypothesis is that the artifacts we collected after the game were much more nuanced than the original artifact list from the survey. Since we manually verified the outputs of LLMs, we discovered that they misclassify culturally similar artifacts and align more towards more "popular" artifacts. For instance, Karwa Chauth and Teej festivals involve women fasting for their husbands' long lives. However, Karwa Chauth is specific to Punjabi culture and has been widely represented in popular culture through Bollywood movies, while Teej is predominantly celebrated in Uttar Pradesh and Rajasthan. This misclassification persists even when utilizing prompt engineering techniques, which include providing explicit location information (Fig 3) and festival-specific clues from the clue list.

6. Discussion & Future Work

With the increase in the use of LLMs for various tasks, there has been an increase in evaluating LLMs for values, knowledge, and biases encoded in them. However, each nation itself is culturally diverse. Our work focused on assessing the LLMs at the geographic subcultural level in India. Past work shows that most models treat users from the same Euro-centric lens and assume knowledge of the Global North as the default, resulting in less representative outputs. The lack of representativeness is usually attributed to the lack of training data



(a) **Original Artifacts**



(b) **Expanded Artifacts**

Figure 4: Overall accuracy of different models across 18 States in India on both the Original Artifacts (Fig 4a) and Expanded Artifacts (Fig 4b). Due to operational constraints, Madhya Pradesh (denoted in grey) was excluded from data collection.

States	**Models**			
	Open Source		*Commercial*	
	Llama 2	**Falcon**	**GPT-4**	**Palm 2**
Andhra Pradesh	0.07	0.14	**0.43**	0.36
Assam	0.18	0	**0.45**	0.36
Bihar	0.50	0.16	0.50	**0.75**
Chhattisgarh	0.11	0	**0.22**	0.11
Delhi	0.30	0	**0.50**	0.40
Gujarat	0.22	0.05	**0.38**	0.33
Haryana	0.42	0.08	0.58	**0.83**
Jharkhand	0.33	0.04	**0.71**	0.57
Karnataka	0.21	0	**0.47**	0.42
Kerala	0.62	0.19	**0.75**	0.69
Maharashtra	0.62	0.06	**0.69**	**0.69**
Odisha	0.25	0	**0.58**	0.16
Punjab	0.50	0.25	**0.65**	0.55
Rajasthan	0.36	0.18	0.63	**0.73**
Tamil Nadu	0.55	0	**0.70**	0.60
Telangana	0.46	0.07	**0.54**	0.46
Uttar Pradesh	0.61	0	**0.69**	0.54
West Bengal	0.40	0.05	**0.55**	0.50

Table 2: The table denotes the overall accuracy of the LLMs on the **Original Artifacts** in DOSA.

from "certain cultures or regions." In this work, we created a novel knowledge dataset of social artifacts using participatory research and analyzed the cultural familiarity of the four most well-known and widely used LLMs. We find that LLMs have variance in their familiarity, but they are not entirely unaware of these artifacts. This aligns with past work that shows that LLMs prefer American cultural values in chat settings (Cao et al., 2023). In multicultural contexts, chatbots violating social norms can lead to communication breakdown (Jurgens et al., 2023). Hence, it's crucial to evaluate why LLMs aren't representing their awareness of social artifacts in their outputs.

Learnings: The conversations and participants' feedback were precious sources of information. All participants agreed that the artifacts in the survey response were important and well-known to the community. However, they also mentioned that many of these artifacts could be considered "popular" in India. The game served as an excellent way of priming our participants, and post-game, they showed excitement in sharing more artifacts from their community. We see this in the significant increase in the new and nuanced artifacts we collected. One of the participants mentioned that "the game was very enjoyable, and it made us remember all these objects and things that are very implicit to us, and we do not necessarily think about them as being different from us." The richness of artifacts collected varied across participants, which raises an important question of "whom to consult?" Next, we also observed that states that have large geographic regions, like Karnataka and Uttar Pradesh, have a significant in-state variance in the names of artifacts and sometimes their uses. We also

observed that certain social artifacts are similar across neighboring states, and this is in concurrence with our survey questions on "What other state do you believe is culturally similar to the one that you identify with?" Unlike human participants, we found that LLMs could lose the logical coreference connections despite having all relevant details in the prompt (refer Fig 3), and adding minimal context like 'this *flower* is' helps the LLMs perform better.

Future Work: This work looks at one dimension of culture - geographic boundaries and takes a step towards introducing how NLP can be made more community-centered. However, there are multiple dimensions, like gender, caste, and race, that shape culture, and LLMs need to be aware of these. Future work should investigate how to evaluate the other dimensions of culture. Given the lack of datasets, especially community-centered datasets, future work should investigate how participatory research can be scaled up to provide more breadth and depth to the datasets created. Creating these datasets would also help preserve knowledge and ensure that LLMs do not lend themselves to cultural erasure.

7. Limitations

Our study is subject to a few important limitations. First, while we cover 18 states, some states and union territories are still missing from the dataset. Second, culture is a highly complex concept shaped by the different identities of individuals, and these intersectionalities of identities lend themselves to the social artifacts considered important by that community. Our work does not systematically recruit from subcultures within each state, making our list of social artifacts incomplete. Third, the language of the survey and the game was English. Since the equivalent name for many objects may not exist in other languages or be unknown to most community members, the language would have limited the responses we got in the survey and impacted the diversity of the participants in both the survey and the game.

8. Ethical Considerations

We use the framework by Bender and Friedman (2018) to discuss the ethical considerations for our work.

Institutional Review: All aspects of this research were reviewed and approved by the Institutional Review Board of a research lab in India.

Curation Rationale: To study the cultural familiarity of LLMs, surveys and a Game with a Purpose (GWAP) were conducted. The researchers did not exclude any artifacts given by the survey takers. The vetting by the participants from GWAP ensured that no harmful data made it to the final dataset.

Language Variety: The participants for both the survey and a Game with a Purpose (GWAP) culturally identify as Indian. The artifacts are transliterated, and hence, the language would be a mix of (the US variant of English) en-US, (the British variant of English) en-GB, and (the Indian variant of English) en-IN. While transcribing, we transliterated the artifacts and clues to a mix of en-US, en-GB, and en-IN. The researchers were the observers and moderators, ensuring no offensive stereotypes were included in the data.

Speaker Demographic: In this version of the study, we do not ask participants to disclose their demographics; this was done to make more participants comfortable in engaging with the questions about their cultural identity and the artifacts they perceive as important.

Speech Situation: The speeches in GWAP were informal, spontaneous, and intended for participants from the same geographical culture.

9. Bibliographical References

NITI Aayog. 2023. National multidimensional poverty index.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463.

Anurag Acharya, Kartik Talamadupula, and Mark A Finlayson. 2020. Towards an atlas of cultural commonsense for machine reasoning. *arXiv preprint arXiv:2009.05664*.

Umair Z Ahmed, Kalika Bali, Monojit Choudhury, and VB Sowmya. 2011. Challenges in designing input method editors for indian lan-guages: The role of word-origin and context. In *Proceedings of the Workshop on Advances in Text Input Methods (WTIM 2011)*, pages 1–9.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2022. Probing pre-trained language models for cross-cultural differences in values. *arXiv preprint arXiv:2203.13722*.

Mohammad Atari, Mona J Xue, Peter S Park, Damián Blasi, and Joseph Henrich. 2023. Which humans?

Agathe Balayn, Gaole He, Andrea Hu, Jie Yang, and Ujwal Gadiraju. 2022. Ready player one! eliciting diverse knowledge using a configurable game. In *Proceedings of the ACM Web Conference 2022*, pages 1709–1719.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Johana Cabrera, M Soledad Loyola, Irene Magaña, and Rodrigo Rojas. 2023. Ethical dilemmas, mental health, artificial intelligence, and llm-based chatbots. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 313–326. Springer.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

Harrison Chase. 2022. LangChain.

Harshita Diddee, Kalika Bali, Monojit Choudhury, and Namrata Mukhija. 2022. The six conundrums of building and deploying language technologies for social good. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)*, pages 12–19.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2023. Bias of ai-generated content: An examination of news produced by large language models. *arXiv preprint arXiv:2309.09825*.

Ronald Fischer, Markus Luczak-Roesch, and Johannes A Karl. 2023. What does chatgpt return about human values? exploring value bias in chatgpt using a descriptive value theory. *arXiv preprint arXiv:2304.03612*.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Kanika Gupta, Monojit Choudhury, and Kalika Bali. 2012. Mining hindi-english transliteration pairs from online hindi lyrics. In *LREC*, pages 2459–2465.

Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovickỳ, Constantin A Rothkopf, Alexander Fraser, and Kristian Kersting. 2022. Speaking multiple languages affects the moral bias of language models. *arXiv preprint arXiv:2211.07733*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*.

Geert Hofstede. 2011. Dimensionalizing cultures: The hofstede model in context. *Online readings in psychology and culture*, 2(1):8.

R. Inglehart, C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, E. Ponarin P. Norris, and B. Puranen et al. (eds.). 2018. World values survey: Round six - country-pooled datafile. madrid, spain & vienna, austria: Jd systems institute & wvsa secretariat. *doi.org/10.14281/18241.8*.

Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the benefits and challenges of deploying conversational ai leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

David Jurgens, Agrima Seth, Jackson Sargent, Athena Aghighi, and Michael Geraci. 2023. Your spouse needs professional help: Determining the contextual appropriateness of messages through modeling social relationships. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10994–11013.

Hadas Kotek, Rikker Dockum, and David Q Sun. 2023. Gender bias and stereotypes in large language models. *arXiv preprint arXiv:2308.14921*.

Virginie Mamadouh. 2020. Writing the world in 301 languages: A political geography of the online encyclopedia wikipedia. *Handbook of the Changing World Language Map*, pages 3801–3824.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: Measuring chatgpt political bias. *Public Choice*, pages 1–21.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, pages 1907–1917.

OpenAI. 2023. Gpt-4 technical report.

Shramay Palta and Rachel Rudinger. 2023. Fork: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural incongruencies in artificial intelligence. *arXiv preprint arXiv:2211.13069*.

Rida Qadri, Renee Shelby, Cynthia L Bennett, and Emily Denton. 2023. Ai's regimes of representation: A community-centered study of text-to-image models in south asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 506–517.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Agrima Seth, Jiyin Cao, Xiaolin Shi, Ron Dotsch, Yozen Liu, and Maarten W Bos. 2023. Cultural differences in friendship network behaviors: A snapchat case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Devinder Pal Singh and Manoj K Sharma. 2009. Unfolding the indian cultural mosaic: a cross-cultural study of four regional cultures. *International Journal of Indian Culture and Business Management*, 2(3):247–267.

Janet Stephenson. 2023. *Culture and Sustainability: Exploring Stability and Transformation with the Cultures Framework*. Springer Nature.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*.

Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. Exploring large language models' cognitive moral development through defining issues test. *arXiv preprint arXiv:2309.13356*.

Vishesh Thakur. 2023. Unveiling gender bias in terms of profession across llms: Analyzing and addressing sociological implications. *arXiv preprint arXiv:2307.09162*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Luis Von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 75–78.

Han Wang, Ming Shan Hee, Md Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2023. Evaluating gpt-3 generated explanations for hateful content moderation. *arXiv preprint arXiv:2305.17680*.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From instructions to intrinsic human values–a survey of alignment goals for big models. *arXiv preprint arXiv:2308.12014*.

A. Appendix

A.1. Survey Questionnaire

Cultural identity is a way of belonging to a social group with the same intrinsic features and characteristics. Culture is a complex concept, and Individuals consider many factors while constructing

their cultural identities: location, sexuality, race, history, nationality, language, religious beliefs, and ethnicity.

In this study, we ask you to define your cultural identity based on the regions or states in India. Regional identity is a major factor in determining one's cultural identity, so we encourage you to consider this when defining your cultural identity.

Q1. What is your cultural identity?

- Andhra Pradesh
- Assam
- Bihar
- Chhattisgarh
- Delhi
- Gujarat
- Haryana
- Jharkhand
- Karnataka
- Kerala
- Madhya Pradesh
- Maharashtra
- Orissa
- Punjab
- Rajasthan
- Tamil Nadu
- Uttar Pradesh
- West Bengal
- Other

Q2. Please list identities that you believe most individuals from your cultural identity would associate with the most or be most familiar with.

- Andhra Pradesh
- Assam
- Bihar
- Chhattisgarh
- Delhi
- Gujarat
- Haryana
- Jharkhand
- Karnataka
- Kerala
- Madhya Pradesh
- Maharashtra
- Orissa
- Punjab
- Rajasthan
- Tamil Nadu
- Uttar Pradesh
- West Bengal
- Other

Section 2

Social Artifacts As members of a culture, we create objects that help us connect and stay associated with the culture. These objects are known and have significance to most people who consider themselves as a part of that culture and serve as a way of identifying themselves with the culture and the people in that culture. These objects are called social artifacts.

Here are some categories of artifacts that can help guide your thinking.

1. Names of animals like elephants hold importance in Kerela.
2. Food and beverages like ghewar hold importance in Rajasthan.
3. Clothing like Mysore silk sarees holds importance in Karnataka.
4. Home-related ornamentations like Kollam in houses in Tamil Nadu.
5. Rituals and customs
6. Names of handicrafts, dance and music forms, or
7. Locations like the name of a particular park, shop, monument, beach, etc.

Q3. Write down the 5 'Social Artifacts' that are commonly known, hold relevance to the cultural identity you indicated in question 1, and, to the best of your knowledge, are known to 'a significant' number of people who share the same cultural identity.

Section 3

Q4. Please list identities that you believe most individuals from your cultural identity would least associate with or be familiar with the least.

- Andhra Pradesh

- Assam
- Bihar
- Chhattisgarh
- Delhi
- Gujarat
- Haryana
- Jharkhand
- Karnataka
- Kerala
- Madhya Pradesh
- Maharashtra
- Orissa
- Punjab
- Rajasthan
- Tamil Nadu
- Uttar Pradesh
- West Bengal
- Other

Q5. What three other cultural identities are you most familiar with or associate with the most?

- Andhra Pradesh
- Assam
- Bihar
- Chhattisgarh
- Delhi
- Gujarat
- Haryana
- Jharkhand
- Karnataka
- Kerala
- Madhya Pradesh
- Maharashtra
- Orissa
- Punjab
- Rajasthan
- Tamil Nadu
- Uttar Pradesh
- West Bengal
- Other

A.2. Survey Data Cleaning

The survey response to the question on social artifacts was free-text. Hence, the data was manually cleaned and consolidated to account for differences in spelling and mixed cases for the same artifact. For example, Bandhani and Bandhej refer to the same textile technique and were treated as the same artifact. We did not discard either of the words but treated them as the same artifact.

A.3. Accuracies at GUESS1 and GUESS2 for Original artifacts

The accuracy@GUESS1 and accuracy@GUESS2 for original artifacts in DOSA are reported in Table 3

A.4. Benchmarking results for the expanded artifacts in DOSA

We replicate the benchmarking methodology in Section 4 and calculate the evaluation metrics 4.2 for the artifacts collected during the game, i.e., the expanded dataset. For overall accuracy, refer to Table 4 and Figure 4b. Accuracy@GUESS1 and GUESS2 are in the table 5

States	Models			
	Open Source		*Commercial*	
	Llama 2	**Falcon**	**GPT-4**	**Palm 2**
Andhra Pradesh	0.07 / 0	0.14 / 0	0.36 / 0.11	0.28 / 0.10
Assam	0.18 / 0	0 / 0	0.36 / 0.14	0.27 / 0.14
Bihar	0.25 / 0.33	0.16 / 0	0.42 / 0.14	0.50 / 0.50
Chhattisgarh	0.11 / 0	0 / 0	0.22 / 0	0.11 / 0
Delhi	0.30 / 0	0 / 0	0.5 / 0	0.30 / 0.14
Gujarat	0.22 / 0	0.05 / 0	0.27 / 0.15	0.33 / 0
Jharkhand	0.33 / 0	0.04 / 0	0.62 / 0.25	0.38 / 0.31
Haryana	0.42 / 0	0 / 0.08	0.58 / 0	0.58 / 0.60
Karnataka	0.21 / 0	0 / 0	0.47 / 0	0.42 / 0
Kerala	0.56 / 0.14	0.19 / 0	0.63 / 0.33	0.63 / 0.16
Maharashtra	0.56 / 0.14	0.06 / 0	0.69 / 0	0.44 / 0.44
Odisha	0.16 / 0.10	0 / 0	0.42 / 0.28	0.16 / 0
Punjab	0.45 / 0.09	0.25 / 0	0.65 / 0	0.55 / 0
Rajasthan	0.36 / 0	0.18 / 0	0.63 / 0	0.63 / 0.25
Tamil Nadu	0.50 / 0.10	0 / 0	0.65 /0.14	0.45 / 0.27
Telangana	0.38 / 0.13	0.07 / 0	0.46 / 0.14	0.38 / 0.13
Uttar Pradesh	0.54 / 0.17	0 / 0	0.38 / 0.25	0.54 / 0.33
West Bengal	0.25 / 0.20	0.05 / 0	0.50 / 0.10	0.35 / 0.23

Table 3: The above table shows guess-wise accuracy on the **Original Artifacts** in DOSA. The first number shows the **accuracy@GUESS1**, and the second denotes the **accuracy@GUESS2**.

States	Models			
	Open Source		*Commercial*	
	Llama 2	**Falcon**	**GPT-4**	**Palm 2**
Andhra Pradesh	0.08	0	**0.39**	0.16
Assam	0.08	0.02	**0.09**	0.02
Bihar	0.17	0	**0.34**	0.17
Chhattisgarh	0	0	0	0
Delhi	NA	NA	NA	NA
Gujarat	0.04	0	0.07	**0.13**
Haryana	0.12	0.07	0	0
Jharkhand	0.34	0	**0.34**	0.08
Karnataka	0.25	0	**0.38**	0.13
Kerala	0.31	0.16	**0.70**	0.24
Maharashtra	0.23	0.06	**0.62**	0.45
Odisha	0.1	0	**0.13**	**0.13**
Punjab	0.12	0	**0.48**	0.28
Rajasthan	0.17	0.17	**0.67**	0.50
Tamil Nadu	0.11	0	**0.35**	0.11
Telangana	0.12	0	**0.12**	0.08
Uttar Pradesh	0.06	0	**0.39**	0.21
West Bengal	0.06	0	**0.56**	0.17

Table 4: The above table shows the overall accuracy of the **Expanded Artifacts** in DOSA. (We could not collect any unique expanded artifacts from the participants of Delhi during our interviews.)

States	Models			
	Open Source		*Commercial*	
	Llama 2	**Falcon**	**GPT-4**	**Palm 2**
Andhra Pradesh	0.08 / 0	0 / 0	0.15 / 0.27	0.15 / 0
Assam	0.07 / 0	0.02 / 0	0.05 / 0.04	0.02 / 0
Bihar	0.17 / 0	0 / 0	0.17 / 0.20	0.17 / 0
Chattisgarh	0 / 0	0 / 0	0 / 0	0 / 0
Delhi	NA	NA	NA	NA
Gujarat	0.03 / 0	0 / 0	0.06 / 0	0.13 / 0
Haryana	0.06 / 0.06	0 / 0	0 / 0	0 / 0
Jharkhand	0.33 / 0	0 / 0	0.33 / 0	0.08 / 0
Karnataka	0.25 / 0	0 / 0	0.31 / 0.09	0.13 / 0
Kerala	0.15 / 0.18	0.15 / 0	0.61 / 0.20	0.23 / 0
Maharashtra	0.22 / 0	0.05 / 0	0.44 / 0.30	0.38 / 0.09
Odisha	0.09 / 0	0 / 0	0.13 / 0	0.13 / 0
Punjab	0.12 / 0	0 / 0	0.40 / 0.13	0.28 / 0
Rajasthan	0.17 / 0	0.17 / 0	0 / 0.67	0.17 / 0.40
Tamil Nadu	0.07 / 0.03	0 / 0	0.20 / 0.17	0.10 / 0
Telangana	0.11 / 0	0 / 0	0.11 / 0	0.07 / 0
Uttar Pradesh	0.05 / 0	0 / 0	0.38 / 0	0.11 / 0.07
West Bengal	0.05 / 0	0 / 0	0.38 / 0.27	0.16 / 0

Table 5: The above table shows guess-wise accuracy on the **Expanded Artifacts**. The first number shows the **accuracy@GUESS1**, and the second denotes the **accuracy@GUESS2**.

State	Artifact	Clues
Punjab	Lohri	this is a Punjabi festival usually celebrated in winters people make bonfires and kites are flown during the day this festival is also known as festival of kites people usually eat peanuts, rewaris etc. during this festival
	sarson ka saag	this is a food item that is usually had in winters It is made using the leaves of the mustard plant
Maharashtra	Lavani	It is an old form of dance This form of dance is usually practiced when the villagers are done harvesting their crops
	Tandalachi bhakri	this is an alternative to roti it is made of rice it is white in color
Assam	Jappi	A headgear usually worn by farmers, fishermen, and tea garden workers. These days, it has been commodified for gifting. It is made of bamboo straws, and the ones that are souvenirs sometimes have designs made of velvet clothes.
	Gamusa	A traditional garment which is usually used to wipe oneself. Traditional ones are white and red in color. The patterns or designs on it are used to distinguish the use of this garment.
Tamil Nadu	Veshti	It is usually a part of men's attire It is a kind of Long cloth
	kolam	It's found in a Tamil house Outside the door area Made out of white powder

Table 6: The above table shows a few of the artifacts and their corresponding clues collected during the participatory research