Evaluating AI-Generated Images of Cultural Artifacts with Community-Informed Rubrics

Nari Johnson[1,2], Deepthi Sudharsan[†1,3], Hamna[†1], Samantha Dalal[1,4], Theo Holroyd[5], Anja Thieme[1], Hoda Heidari[2], Daniela Massiceti[1], Jennifer Wortman Vaughan[1], Cecily Morrison[1]
[1]Microsoft Research, [2]Carnegie Mellon University, [3]RiskSpan,
[4]Princeton University, [5]The Stephen Perse Foundation

Abstract

Recent efforts to automate and quantify generative AI evaluation can be in tension with the ability for measurement instruments to capture the expertise and perspectives of communities impacted by AI. In this paper, we explore how to involve communities in drafting *evaluation rubrics* that can be used to score AI images of cultural content. Specifically, we systematize the concept of "culturally appropriate" depictions of cultural content (*i.e.,* culturally significant objects) through case studies with three communities: blind and low vision individuals residing in the UK and residents of two distinct Indian geographic states. Our systematized concepts reflect community members' lived experiences and desires for cultural representation, demonstrating the value of community involvement in defining valid measures. We explore how these systematized concepts can be operationalized into automated measurement instruments that could be applied across contexts using a multimodal LLM-as-a-judge approach. We point to timely opportunities for research to advance methods that bring community expertise into the sociotechnical measurement of generative AI systems.

Note: This paper provides a short version of our research that was presented at the ACL 2026 Workshop on Evaluating Evaluations. A longer version was published in ACM FAccT 2026, and can be accessed online at (this link).

1 Introduction

People across the world have adopted generative AI tools to automate the creation of images for design and ideation, marketing and advertising, illustration, and beyond (Jiang et al., 2023; Dehouche and Dehouche, 2023; Gillespie, 2024). But not all cultures are rendered accurately or appropriately. HCI and NLP researchers have documented the many ways in which state-of-the-art generative AI systems systematically underperform at depicting historically marginalized cultures, including replicating historical biases in media (Gillespie, 2024; Mack et al., 2023), reinforcing stereotypes (Bianchi et al., 2023; Jha et al., 2024; Hall et al., 2024; Gautam et al., 2024), and contributing to cultural erasure (Ghosh et al., 2024; Johnson et al., 2025; Magomere et al., 2025; Qadri et al., 2025).

Effective model evaluation is necessary to make these failures visible, and measure progress toward addressing them (Weidinger et al., 2023; Wallach et al., 2025; Harvey et al., 2025). Today, many existing approaches to generative AI evaluation apply off-the-shelf benchmarks and metrics (Eriksson et al., 2025; Kapania et al., 2025; Harvey et al., 2025). However, an emerging line of research suggests that these existing approaches to evaluation suffer from significant validity issues (Wallach et al., 2025; Coston et al., 2023), and in particular, break down when applied in marginalized, low-data contexts (Kreiss et al., 2022; Johnson et al., 2025; Hada et al., 2024).

An emerging body of human-centered literature has explored ways to involve humans that hold relevant "lived-experience expertise" (*e.g.,* due to their knowledge, standpoint, or cultural identity) (Matias and Price, 2025) in the process of evaluating generative AI. This ranges from efforts that crowdsource culturally specific datasets (Orife et al., 2020; Seth et al., 2024; Magomere et al., 2025) to those that invite participants to evaluate model outputs (Mack et al., 2023; Aroyo et al., 2023; Hall et al., 2024; Qadri et al., 2025). At the same time, there is a rapidly growing literature on ways to develop practical, repeatable, and automatable evaluation practices by using MLLM-as-a-judge evaluations, which delegate evaluative judgments to multimodal language models (MLLMs) in lieu of hu-

[†]Second authors, equal contribution. Work done at MSR.

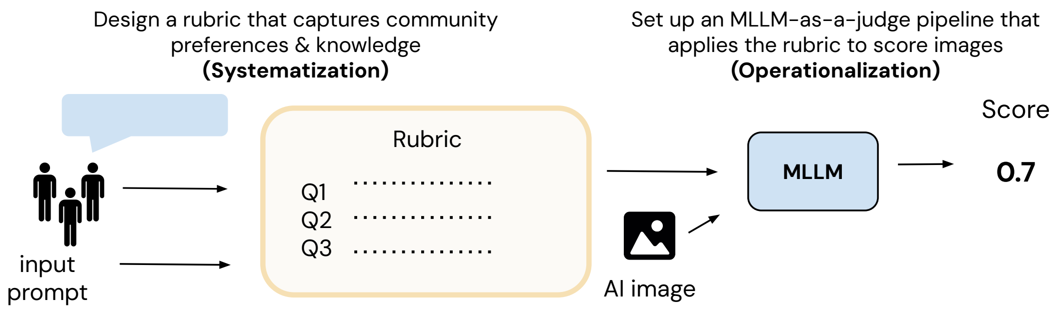


Figure 1: **Scaffolding community engagement to develop community-centered measures of cultural representation**. Given an input prompt (*e.g.,* "*a photo of a guide cane*"), we invited community members to participate in designing a rubric that captures their expertise and preferences for each cultural artifact (systematization). Our research team then explored the use of this rubric within an automated multimodal LLM-as-a-judge pipeline (operationalization).

man judges (Shankar et al., 2024; Szymanski et al., 2024b). This paper aims to connect these two literatures by engaging people with lived-experience expertise in developing rubrics that can be used within the MLLM-as-a-judge paradigm to create automated measurement instruments (Figure 1).

Recent work by Wallach et al. (2025) calls for AI researchers to reimagine their approach to measurement by adopting an established measurement framework from the social sciences (Adcock and Collier, 2001). The framework breaks the process of measurement into three concrete steps: (1) *systematizing* an abstract concept to be measured into a concrete definition; (2) *operationalizing* the systematized concept into a measurement instrument; and (3) *applying* the measurement instrument to produce measurements. The authors point to the opportunity for participants with different forms of expertise to be included in the systematization process, but stop short of offering practical guidance on how to facilitate such engagements. In this work, we ask: how can we scaffold community engagement in the systematization process to develop community-informed measurement instruments?

To answer this question, we facilitate research workshops with members of three communities: blind and low vision individuals residing in the UK and residents of two distinct South Indian states, Kerala and Tamil Nadu. We follow past HCI studies (Bergman et al., 2024; Hamna et al., 2025a; Matias and Price, 2025) in using the term "community" to refer to a group of individuals who hold a shared cultural identity, recognizing the shared "lived-experience expertise that individuals hold that extends beyond their formal training or credentials", such as the embodied knowledge that is held by a blind guide cane user. Our goal is to identify ways for non-technical people (*e.g.,* people who are not AI developers) to contribute their lived-experience expertise in shaping AI evaluation design. In each context, we develop community-informed systematized concepts of *cultural appropriateness* in depicting cultural artifacts (Newmark, 1988). We view cultural appropriateness loosely as the degree to which representations of artifacts align with and respect the values, norms, and knowledge systems of the relevant community. In studying culturally appropriate depictions, our goal is to capture community members' understandings of how each artifact should—and critically, should not—be represented in AI media (Qadri et al., 2023; Mack et al., 2023; Qadri et al., 2025). We structure our inquiry around the following research questions:

- **RQ1 (Systematization):** How can practitioners work with community members to systematize the cultural appropriateness of several important cultural artifacts? How do the rubrics elicited through a community-centered approach differ from those generated by LLMs?

- **RQ2 (Operationalization):** How might practitioners automate the application of community-informed rubrics to score new AI images?

We contribute a rich empirical account of how practitioners can engage marginalized communities in the systematization process. We highlight how community members' expertise, lived and embodied experiences with each artifact, and subjective preferences for representation shape our systematized concepts, demonstrating the potential value of involving communities in measurement. We explore whether rubric application can be automated using an MLLM judge, surface several limitations of MLLM judges, and contribute methodological guidance for how practitioners can assess the feasibility of automating measurement. We conclude by sharing learnings and opportunities for future research to bring community expertise into the design of AI evaluation metrics.

2 Related Work

Cultural Representation and AI-Generated Media. AI-generated images have the power to shape how communities are understood and perceived by others, as well as how they perceive themselves (Qadri et al., 2023; Hall, 1997). Marginalized communities have used AI tools to create visual media that aligns with their goals for representation (Das et al., 2024; Huh et al., 2023; Adnin and Das, 2024; He et al., 2025; Hamna et al., 2025b). However, these same AI tools have been shown to reproduce normative identities and narratives that contribute to the erasure of already marginalized communities (Bianchi et al., 2023; Gillespie, 2024; Mack et al., 2024), resulting in representational harm. In this paper, we focus on the issue of appropriate and desired cultural representation—an often contested concept (Chasalow and Levy, 2021). We adopt a broad conceptualization of culture as based on shared identity, recognizing that individuals can inhabit multiple cultural identities simultaneously (Zhou et al., 2025). Culture may be grounded in place (*e.g.,* shared nationality (Adilazuarda et al., 2024)), or in other identity aspects including race or ethnicity (Egede, 2025), disability (Mack et al., 2024), sexuality (Taylor et al., 2024), profession (Tseng et al., 2025), or relational roles (Viswanathan et al., 2025).

In this paper, we focus on designing quantitative measures that reflect participants' lived understandings of their culture. Prior work in AI and culture has emphasized the need for methodologies that "translate qualitative insights into algorithmic interventions" (Biega et al., 2025). Our work responds to this call by bringing qualitative methods into the design of quantitative metrics by engaging communities in the design of rubrics that can be applied with MLLM-as-a-judge systems.

Measuring the Performance of Generative Text-to-Image Models. Our work contributes to a growing body of interdisciplinary scholarship focused on developing and critiquing performance measures used to evaluate generative AI systems. Given the speed at which the AI industry is moving, there is a need for evaluation practices to be repeatable, automatable, and applicable across different models, datasets, and settings (Shankar et al., 2024; Weidinger et al., 2023; Harvey et al., 2025). It is increasingly common to automate evaluations using model-mediated evaluation approaches, where an auxiliary, pretrained model's internal representations or judgments are used to stand in for human evaluative judgments as a form of "ground truth" (Kapania et al., 2025). Drawing on the affordances for scale of LLM-as-a-judge pipelines, practitioners have increasingly turned to multimodal language models (MLLMs) as judges for evaluating generated images (Hu et al., 2023; Zhang et al., 2023; Lin et al., 2024; Chen et al., 2025). In these pipelines, a multimodal model scores each image using a rubric, which in some cases is generated by another model. Such approaches, however, introduce a fundamental dependency: the quality of the measurements becomes limited by the capabilities and biases of the auxiliary models (Magomere et al., 2025; Kreiss et al., 2022; Seth et al., 2024; Kapania et al., 2025; Massiceti et al., 2024). Since these models are trained on historical datasets that may lack diversity or contain societal biases, they can systematically undervalue or misrepresent content from marginalized communities. We ask whether this limitation can be overcome by the design of more thoughtful rubrics.

3 Methods

Selecting communities and cultural artifacts. We collaborated with members of three distinct communities: blind and low-vision (BLV) people located in the United Kingdom, and current and former residents of two states in South India: Tamil Nadu and Kerala. We chose these specific communities because members of the research team belong to them. This close researcher–community collaboration supported activities such as artifact selection, participant recruitment, and interpretation of our findings.

For the BLV community, we focused our study on a single region (the UK) where our authorship team was based. We chose to geographically situate our study in the UK because the material culture of blindness is shaped by local factors, such as the public funding of assistive technology use

Figure 2: **Selected culturally significant artifacts.** From right to left: (1) With the blind and low vision community, we selected a *guide cane* (a mobility aid that is held diagonally across one's body) and a *braille notetaker* (an electronic device that can be used to read and write notes in tactile braille). (2) With residents of Tamil Nadu, we selected *Pallanguzhi* (a two-player mancala game where players compete to collect cowry shells or seeds) and *Mridangam* (a percussion instrument widely used in South Asian classical music). (3) With residents of Kerala, we selected a *Kasavu saree* (a handwoven saree from Kerala, known for its off-white body with a gold border), and *Chundan Vallam* (a traditional boat from Kerala with a raised prow commonly used in festival races).

in schools (Connell, 2008). We also conducted two separate engagements with residents of two Indian states: Tamil Nadu and Kerala. Cultural studies scholars have discussed how India's 28 region-states function not merely as administrative regions, but also each carry their own unique histories and cultural practices that play an important role in creating a shared cultural identity (Tambiah, 1967; Singh and Sharma, 2009). We follow past work (Seth et al., 2024) and center community engagements on cultural artifacts that are *unique to each state*, and do not have a direct equivalent in neighboring geographic states. Each engagement was facilitated by a co-author who was from each state.

Guided by the members of our research team who belong to each community, we selected salient cultural artifacts (Newmark, 1988). These artifacts are not intended to represent the richness and breadth of each community's material culture (Zhou et al., 2025), but to serve as a starting point for our exploration of rubric design. For the BLV community, we selected two assistive technologies: a *guide cane*, a mobility device that helps its user detect obstacles and navigate their surroundings, and a *braille notetaker*, an electronic device used to write and read notes in tactile braille. The particular technologies were chosen because of their wide use by and cultural significance to the BLV community in the UK, and are pictured in Figure 2. For the two Indian communities, we selected objects regarded as meaningful to their state's cultural identity and likely to be known by residents. From Kerala, we selected (1) the *Kasavu saree*, a traditional white and gold garment, and (2) the *Chundan Vallam*, a snake boat used in boat racing festivals. From Tamil Nadu, we selected (1) the *Pallanguzhi*, a traditional two-player mancala board game, and (2) the *Mridangam*, a percussion instrument that is widely used in South Indian classical music. These objects, pictured in Figure 2, carry historic and symbolic value for their roles in each state's unique cultural ceremonies, festivals, and traditions (Newmark, 1988; Yao et al., 2024).

Eliciting culturally appropriate artifact depictions in workshops. To create the rubrics, we conducted a series of synchronous, hour-long online workshops with community members. Workshops were facilitated by members of our research team, and took place between December 2024 to April 2025. To recruit participants, we used a purposive sampling approach (Palinkas et al., 2013), drawing on the research team's existing networks across each community. Our final sample included 10 BLV participants in the UK, 9 participants from Tamil Nadu, and 8 participants from Kerala. Participants were compensated either £75 (UK) or 500 INR (India) based on their location, and all interviews were conducted in English. All workshop studies were approved by our institution's ethics review board. We detail the structure of the workshops, including activity guides, participant demographics, and steps we took to ensure the workshops were accessible for members of the BLV community in Appendix D.2.1.

For each artifact, we curated a small set of AI-generated images and real photographs to facilitate discussions of cultural representation. We generated images using two state-of-the-art text-to-image models at the time of the study, Stable Diffusion 3 (Stability AI, 2024a) and DALL·E 3 (OpenAI, 2024), prompting each model with a simple template (*e.g.,* "A photo of a {artifact name}"). For each artifact, we generated an initial pool of images and then grouped images by shared visual characteristics, sampling across groups to create a manageable but meaningfully diverse set for participants to review (Mack et al., 2023). The final set of images used in the workshops is included in Appendix D.

For each artifact, we conducted two study activities. First, after presenting each image (by screen-sharing and/or presenting an alt-text description for BLV participants), the facilitator asked participants to share if they felt that the image could or could not be shown to the general public to represent the artifact and why. The facilitator then prompted each participant to elaborate on what made a particular image a good, bad, offensive, or incorrect depiction of the artifact.

Second, we invited participants to reflect on what they had seen so far to create a list of the most important criteria for a culturally appropriate portrayal of the object and provide reasons for each of their responses. This activity encouraged participants to articulate concrete visual criteria that shaped their decisions. The study facilitator encouraged participants to reflect on the degree of acceptable variability for objects and how visual features should be prioritized.

Systematizing cultural representation into rubrics. We used the data collected from workshops to develop systematized concepts of "cultural appropriateness" for each artifact. These took the form of simple rubrics made up of binary yes/no criteria that community members identified as integral to an accurate and appropriate representation of the artifact; see Figure 3 for an example of the rubric produced for a guide cane.

To distill criteria for the rubrics, we first organized participants' statements from both activities into two categories: (1) *criteria*, or the concrete visual features or physical characteristics that shaped participants' decision-making, and (2) *justifications* that participants provided for their decisions. In deciding which criteria to include in the rubrics, we made two key decisions. First, we excluded criteria that were highly contested across participants. We also excluded criteria that participants consistently described as less essential to depicting the essence of each artifact. Consequently, rather than offering a complete description of each artifact, our rubrics focus on the core features that participants agreed were essential to a culturally appropriate depiction.

Finally, we grouped the criteria into higher-level themes that categorize the values motivating the inclusion of different criteria. These themes reflect community-specific understandings of what constitutes (in)appropriate cultural representation. For example, BLV community members emphasized that assistive technologies should be usable and accessible to blind users, clarifying why particular visual features were prioritized. We structured each rubric by these themes.

Using the rubric, we say that an image is "culturally appropriate" only if it satisfies all of the rubric criteria, receiving a positive label of 1. Otherwise, we say that an image is culturally inappropriate (label of 0). While these final binary cultural appropriateness labels do not capture the subjective and contested nature of cultural representation, we adopt this aggregation function for its simplicity to calculate, and its direct comparability with participants' judgments.

As a point of comparison, we also generated rubrics automatically, in line with common AI industry practices (Szymanski et al., 2024a). Specifically, we prompted GPT-4o to produce evaluation criteria describing the "most important visual characteristics that should be present or absent" in culturally appropriate depictions of each artifact (details in Appendix C.1.2). To compare our community-informed rubrics with the automatically generated rubrics, the first author worked with community members on the research team to annotate each rubric, with the aim of identifying overlaps, divergences, and omissions relative to both our community-informed rubrics and our own knowledge of the artifacts.

Operationalizing the rubrics through MLLM-as-a-judge. To operationalize our systematized concepts, we *explore* whether our community-elicited rubrics can be automatically evaluated using an MLLM-as-a-judge pipeline. We adopt a simple system prompt, following past work (Zheng et al., 2023; Chen et al., 2025), that instructs the MLLM to "*determine whether the provided image meets each criterion*". We adopt GPT 4-o (OpenAI

Reference photo

Evaluation Rubric

Themes	Criteria
THEME 1. *"The object needs to be functional as an assistive technology, and usable by someone who is blind."*	C1. No deformed canes. C2. No curved (crooked) handles. C3. The cane must be shaped like a long (5 foot) stick. C4. The body must have sections that are a white color. C5. There must be a tip at the bottom of the cane.
THEME 2. *"The object in the image should not be confused with other, more hegemonic objects, such as objects that are used predominantly by people who are sighted."*	C1. No wooden walking sticks. C2. No decorative striped patterns (candy canes).

Figure 3: **A rubric to score images of a guide cane, designed with BLV community members.** Criteria that correspond to visual features in images are organized under two themes that describe participants' desires for cultural representation.

and others, 2024) as our judge model for its demonstrated performance on MLLM-as-a-judge tasks (Zhang et al., 2023) and report all results averaged over five random seeds. We provide additional details in Appendix C.3.3.

To validate the operationalization—that is, whether the MLLM-as-a-judge measurement instrument captures the systematized concept—we compare judgments of the rubric criteria from an MLLM judge and human annotators on a new dataset of AI-generated images. We create a dataset containing 50 images of each artifact ("*a photo of a guide cane*") generated from 5 different models (DALL·E 3, Stable Diffusion 3 Medium, Stable Diffusion 3.5 Large, GPT Image-1, and Flux.1 Dev). We include image generation details in Appendix C.3.1. These images depict the types of representations and errors that state-of-the-art models produce today. For each image, a member of our research team manually annotated whether each rubric criterion is met, which we compared against annotations assigned by the MLLM judge.

4 Results

In this section, we present the rubrics used to systematize the concept of "culturally appropriate" representation for each cultural artifact. The rubric for a guide cane is shown in Figure 3; the remainder can be found in Appendix A. For each community, we first identified high-level themes that capture key dimensions of how participants evaluated cultural representation. We used these themes to organize a set of criteria that correspond to concrete, observable visual features.

Themes and criteria to assess cultural representation. Across all three communities, participants consistently evaluated images along two core dimensions: *functionality*, whether an artifact could plausibly serve its intended purpose, and *recognizability*, whether the depiction preserves the key features that distinguish the artifact from related, but culturally distinct objects. Within the BLV context, participants drew on their embodied experiences as users of each assistive technology to assess whether a depiction would be able to serve its intended function, *e.g.,* by imagining how one might hold a pictured cane or whether depictions of braille could be read by touch. Participants repeatedly pointed out when generated images "looked like" other recognizable objects that did not serve the intended purpose of the assistive technology, such as a generated image of a guide cane that resembled a walking stick, a distinct mobility aid that differs in its function, material, and handle shape. Community members felt that these "confused" depictions were not only misleading, but were disappointing in that they failed to communicate the unique material culture of blindness.

We found many key similarities in the higher-level themes that describe how residents of Kerala and Tamil Nadu evaluated their cultural artifacts. Participants from Tamil Nadu emphasized evaluat-

Table 1: **Human versus MLLM application of community-informed rubrics.** Using 50 generated images per artifact, we report (i) the proportion of images labeled as culturally appropriate by humans versus an MLLM, and (ii) agreement between MLLM and human labels. Columns further disaggregate agreement by images the human labeled as appropriate versus inappropriate.

Artifact	**Human (% Appropriate)**	**MLLM (% Appropriate)**	**Agreement** Overall	**Agreement** Appropriate images	**Agreement** Inappropriate images
Guide cane	0.40	0.44	0.84	0.84	0.83
Braille notetaker	0.08	0.20	0.82	0.65	0.83
Pallanguzhi	0.18	0.12	0.78	0.22	0.90
Mridangam	0.10	0.21	0.84	0.76	0.85
Kasavu saree	0.12	0.21	0.88	0.87	0.88
Chundan Vallam	0.00	0.17	0.83	N/A	0.83

ing the object's functionality by assessing whether the artifact pictured could be used for its intended purpose, such as how they might set up and play Pallanguzhi, the board game, given the arrangement of pits on a specific board. Participants from Kerala highlighted that generated images failed to capture what they felt were the most culturally distinct features of each artifact, such as the characteristic beak-like tip of the Chundan Vallam racing boat, or the golden hand-woven border that distinguishes a Kasavu saree from sarees emblematic of other regions ("*if it's not white with a gold border, then it's just a saree*").

Assessing the alignment of our rubrics with community preferences. We interrogate the validity of our rubrics by examining whether they align with preferences expressed by community members during workshops. Specifically, we compare (a) labels of cultural appropriateness obtained by having a member of our research team manually evaluate the rubric criteria to (b) participants' judgments of whether or not images were appropriate to be shown from the first workshop activity. We aggregate judgments across workshop participants by taking the majority judgment. We find that for over 80% of images, the label produced by manually applying the rubric matches the majority participant judgment (see Appendix Table 4). Note that we may not expect perfect alignment because of the inherently plural and contested nature of cultural representation, which lacks a singular ground truth. In our own workshops, participants' judgments of cultural appropriateness disagreed with each other for 33% of the workshop images.

Comparing our community-informed rubrics with LLM-generated rubrics. We observe several differences when examining how our rubrics differ from those generated by an off-the-shelf LLM. Many of the LLM-generated rubric criteria accurately capture surface-level properties of the artifacts, such as noting the "distinct golden border" of a Kasavu saree or describing a guide cane as "a long, slender stick" (Appendix C.1.2). However, several LLM-generated rubrics include criteria that are inaccurate, reflecting a fundamental misunderstanding of the objects. For example, the rubric for braille notetaker requires that the depiction "resemble an electronic notebook" and "include a display for visual feedback", two features that are inaccessible to blind users and uncharacteristic of braille notetakers. More generally, most LLM-generated rubrics omitted or underspecified features that community members viewed as essential for functionality or recognizability. The Mridangam rubric, for example, omitted the drum's characteristic black circular membrane that is necessary to its tone. These omissions were not uniform across artifacts: some LLM-generated rubrics, such as that for the Kasavu saree, overlapped substantially with community rubrics, whereas others, most notably the braille notetaker, diverged sharply (see Appendix C.1.2 for further discussion). This suggests that community participation is essential for artifacts or concepts that are at present poorly captured by LLMs.

Operationalizing MLLM-as-a-judge metrics. To assess the feasibility of automating rubric application, we compare MLLM-generated annotations to human annotations for each rubric criterion. Across artifacts, the MLLM's labels of cultural appropriateness agree with human judgments for 78–88% of the images in our dataset (fourth column in Table 1), with a consistent tendency to over-predict appropriateness for five of the six artifacts. We disaggregate agreement by images labeled as appropriate versus inappropriate by humans, and reveal two types of MLLM errors: low agreement on inappropriate images indicates the MLLM's failure to recognize violated criteria, while low agreement on appropriate images (which we see for Pallanguzhi and braille notetakers) may suggest that the MLLM lacks the cultural understanding needed to recognize valid depictions. Each has different implications for how practitioners might revise their rubric criteria to improve judge performance, as the best steps forward (*e.g.,* relaxing or clarifying criteria in cases of false negatives, or strengthening or expanding criteria in cases of false positives) will depend on the type of error being made. We include additional analyses that break down agreement for each individual rubric criterion in Appendix C.3.4.

5 Discussion

This paper centers the perspectives of impacted communities in designing quantitative evaluations of cultural representation in AI-generated images. We provide an initial set of results to show how community expertise and perspectives can be elicited and encoded by *systematizing* criteria for cultural representation in the form of rubrics, and *operationalized* by applying those rubrics with an MLLM-as-a-judge. We hope that researchers and practitioners can extend on the methods presented to invite community members to participate in the design of evaluation metrics. In what follows, we share several lessons learned and open research directions surfaced by this exploratory work. We include an extended discussion of the limitations of our study in Appendix B.

Examining the differences between community-elicited versus LLM-generated rubrics reveals the value of inviting community participation in measurement design. Rather than providing a surface-level description of the visual attributes of each artifact, participants' descriptions spoke to how artifacts are used and experienced in practice. In contrast, many LLM-generated rubrics failed to capture these salient features and in some cases reflected confused or inaccurate understandings. This finding extends those from past work, which describe how LLM-generated rubrics are often overly vague or underspecified (Szymanski et al., 2024a), to reveal how these rubrics may reflect deeper cultural misalignment. Taken together, these results suggest that community participation in defining evaluation criteria can help bridge epistemic gaps in large pretrained models, particularly for low-data cultural artifacts (Massiceti et al., 2024; Hall et al., 2025), improving the validity of the resulting measurement instruments. Beyond asking community members to contribute their knowledge, community participation also enables communities to express their more normative desires and preferences for cultural representation.

Our study points to several promising directions for future work. Iteratively refining rubric criteria to improve their legibility to automated judges is a critical next step for future work and a limitation of this current study that aimed to focus on systematization. Addressing these questions requires grappling with both technical challenges in the design of MLLM-as-a-judge pipelines (Li et al., 2025; Gebreegziabher et al., 2025), and human-centered challenges in thoughtfully engaging community members in a highly specialized and technical measurement design process (Suresh et al., 2023). For example, how can community participants get the feedback needed from the AI model to make useful modifications to the rubric?

While our findings illustrate the potential benefits of human-centered approaches to measurement through close collaborations with three communities and just a small set of artifacts, many open questions remain. In real-world settings, foundation model providers must prioritize engagement across thousands of communities and a wide range of content (Weidinger et al., 2023; Young et al., 2024; Young, 2025), a scale at which both our methods and related qualitative approaches to identifying representational harms (*e.g.,* (Mack et al., 2023; Qadri et al., 2025)) are not feasible to apply. Our findings highlight key considerations for identifying where community participation is most valuable in measurement design. We found that LLM-generated rubrics may serve as a reasonable starting point for some cultural artifacts, but not others. Practitioners may consider inviting community members to assess the face validity of LLM-generated rubrics as a lightweight check. If the rubrics diverge significantly from community members' understandings of the content being evaluated, this is a signal that a deeper engagement (using methods like ours) may be warranted. Practitioners might also consider prioritizing community engagements around content with strong normative stakes and contested meanings, where it is important for community members to exercise authority in shaping how they are represented (Qadri et al., 2025).

References

Robert Adcock and David Collier. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3):529–546.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey.

Rudaiba Adnin and Maitraye Das. 2024. "i look at it as the king of knowledge": How blind people use and understand generative ai tools. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '24, New York, NY, USA. Association for Computing Machinery.

Sherry R. Arnstein. 1969. A ladder of citizen participation. *Journal of the American Institute of Planners*, 35(4):216–224.

Lora Aroyo, Alex S. Taylor, Mark Díaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: diversity in conversational ai evaluation for safety. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Cynthia L. Bennett, Erin Brady, and Stacy M. Branham. 2018. Interdependence as a frame for assistive technology research and design. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '18, page 161–173, New York, NY, USA. Association for Computing Machinery.

Cynthia L. Bennett, Shaun K. Kane, and Christina N. Harrington. 2025. Toward community-led evaluations of text-to-image ai representations of disability, health, and accessibility. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '25, page 256–270, New York, NY, USA. Association for Computing Machinery.

Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. 2024. Stela: a community-centred approach to norm elicitation for ai alignment. *Scientific Reports*, 14(1):6616.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1493–1504, New York, NY, USA. Association for Computing Machinery.

Asia Biega, Georgina Born, Fernando Diaz, Mary L. Gray, and Rida Qadri. 2025. Towards a Multidisciplinary Vision for Culturally Inclusive Generative AI (Dagstuhl Seminar 25022). *Dagstuhl Reports*, 15(1):33–49.

Black Forest Labs. 2024. Flux. `https://github.com/black-forest-labs/flux`.

Stacy M. Branham and Shaun K. Kane. 2015. The invisible work of accessibility: How blind employees manage accessibility in mixed-ability workplaces. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '15, page 163–171, New York, NY, USA. Association for Computing Machinery.

Kyla Chasalow and Karen Levy. 2021. Representativeness in statistics, politics, and machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 77–89.

Jiahui Chen, Candace Ross, Reyhane Askari-Hemmat, Koustuv Sinha, Melissa Hall, Michal Drozdzal, and Adriana Romero-Soriano. 2025. Multi-modal language models as text-to-image model evaluators.

Tim Connell. 2008. The challenge of assistive technology and braille literacy. [Online; accessed 6-September-2025].

Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. 2023. A validity perspective on evaluating the justified use of data-driven decision-making algorithms. In *2023 IEEE conference on secure and trustworthy machine learning (SaTML)*, pages 690–704. IEEE.

Samantha Dalal, Siobhan Mackenzie Hall, and Nari Johnson. 2024. Provocation: Who benefits from "inclusion" in generative ai?

Maitraye Das, Alexander J Fiannaca, Meredith Ringel Morris, Shaun K Kane, and Cynthia L Bennett. 2024. From provenance to aberrations: Image creator and screen reader user perspectives on alt text for ai-generated images. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Maitraye Das, Darren Gergle, and Anne Marie Piper. 2019. "it doesn't win you friends": Understanding accessibility in collaborative writing for people with vision impairments. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Nassim Dehouche and Kullathida Dehouche. 2023. What's in a text-to-image prompt? the potential of stable diffusion in visual arts education. *Heliyon*, 9(6):e16757.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice.

Sunipa Dev, Vinodkumar Prabhakaran, Rutledge Chin Feman, Aida Davani, Remi Denton, Charu Kalia, Piyawat L Kumjorn, Madhurima Maji, Rida Qadri, Negar Rostamzadeh, Renee Shelby, Romina Stella, Hayk Stepanyan, Erin van Liemt, Aishwarya Verma, Oscar Wahltinez, Edem Wornyo, Andrew Zaldivar, and Saška Mojsilović. 2026. A unified framework to quantify cultural intelligence of ai.

Lisa Egede. 2025. *Exploring Black Communities' Perceptions and Design Approaches for Building Culturally Tailored AI Systems*, page 72–76. Association for Computing Machinery, New York, NY, USA.

Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. 2025. Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785.

Sanjana Gautam, Pranav Narayanan Venkit, and Sourojit Ghosh. 2024. From melting pots to misrepresentations: Exploring harms in Generative AI. *arXiv preprint arXiv:2403.10776*.

Simret Araya Gebreegziabher, Charles Chiang, Zichu Wang, Zahra Ashktorab, Michelle Brachman, Werner Geyer, Toby Jia-Jun Li, and Diego Gómez-Zará. 2025. Metricmate: An interactive tool for generating evaluation criteria for llm-as-a-judge workflow. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, CHIWORK '25, New York, NY, USA. Association for Computing Machinery.

Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson, and Aylin Caliskan. 2024. Do generative ai models output harm while representing non-western cultures: Evidence from a community-centered approach. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):476–489.

Tarleton Gillespie. 2024. Generative ai and the politics of visibility. *Big Data & Society*, 11(2):20539517241252131.

Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna Wallach, Zhiwei Steven Wu, and Alexandra Chouldechova. 2025. Validating llm-as-a-judge systems under rating indeterminacy.

Kanika Gupta, Monojit Choudhury, and Kalika Bali. 2012. Mining Hindi-English transliteration pairs from online Hindi lyrics. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2459–2465, Istanbul, Turkey. European Language Resources Association (ELRA).

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation? *arXiv preprint arXiv:2309.07462*.

Melissa Hall, Samuel J. Bell, Candace Ross, Adina Williams, Michal Drozdzal, and Adriana Romero Soriano. 2024. Towards geographic inclusion in the evaluation of text-to-image models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 585–601, New York, NY, USA. Association for Computing Machinery.

Siobhan Mackenzie Hall, Samantha Dalal, Raesetje Sefala, Foutse Yuehgoh, Aisha Alaagib, Imane Hamzaoui, Shu Ishida, Jabez Magomere, Lauren Crais, Aya Salama, and 1 others. 2025. The human labour of data work: Capturing cultural diversity through world wide dishes. *arXiv preprint arXiv:2502.05961*.

Stuart Hall, editor. 1997. *Representation: Cultural Representations and Signifying Practices*. Sage Publications, London.

Hamna, Gayatri Bhat, Sourabrata Mukherjee, Faisal Lalani, Evan Hadfield, Divya Siddarth, Kalika Bali, and Sunayana Sitaram. 2025a. Building benchmarks from the ground up: Community-centered evaluation of llms in healthcare chatbot settings.

Hamna, Deepthi Sudharsan, Agrima Seth, Ritvik Budhiraja, Deepika Khullar, Vyshak Jain, Kalika Bali, Aditya Vashistha, and Sameer Segal. 2025b. Kahani: Culturally-nuanced visual storytelling tool for non-western cultures. In *Proceedings of the 2025 ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, COMPASS '25, page 379–400, New York, NY, USA. Association for Computing Machinery.

Emma Harvey, Emily Sheng, Su Lin Blodgett, Alexandra Chouldechova, Jean Garcia-Gathright, Alexandra Olteanu, and Hanna Wallach. 2025. Understanding and meeting practitioner needs when measuring representational harms caused by llm-based systems.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.

Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. 2025. Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering.

Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. Genassist: Making image generation accessible. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–17.

Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K Reddy, and Sunipa Dev. 2024. Visage: A global-scale analysis of visual stereotypes in text-to-image generation. *arXiv preprint arXiv:2401.06310*.

Harry H. Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. 2023. Ai art and its impact on artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 363–374, New York, NY, USA. Association for Computing Machinery.

Nari Johnson, Hamna Abid, Deepthi Sudharsan, Theo Holroyd, Samantha Dalal, Siobhan Mackenzie Hall, Jennifer Wortman Vaughan, Daniela Massiceti, and Cecily Morrison. 2025. Position: To make text-to-image models that work for marginalized communities, we need new measurement practices for the long tail.

Shivani Kapania, Stephanie Ballard, Alex Kessler, and Jennifer Wortman Vaughan. 2025. Examining the expanding role of synthetic data throughout the ai development pipeline. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. *arXiv preprint arXiv:2205.10646*.

Neha Kumar, Naveena Karusala, Azra Ismail, Marisol Wong-Villacres, and Aditya Vishwanath. 2019. Engaging feminist solidarity for comparative research, design, and practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation.

Kelly Mack, Rai Ching Ling Hsu, Andrés Monroy-Hernández, Brian A. Smith, and Fannie Liu. 2023. Towards inclusive avatars: Disability representation in avatar platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Kelly Avery Mack, Rida Qadri, Remi Denton, Shaun K Kane, and Cynthia L Bennett. 2024. "they only care to show us the wheelchair": disability representation in text-to-image ai models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

Jabez Magomere, Shu Ishida, Tejumade Afonja, Aya Salama, Daniel Kochin, Yuehgoh Foutse, Imane Hamzaoui, Raesetje Sefala, Aisha Alaagib, Samantha Dalal, and 1 others. 2025. The world wide recipe: A community-centred framework for fine-grained data collection and regional bias operationalisation. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 246–282.

Daniela Massiceti, Camilla Longden, Agnieszka Slowik, Samuel Wills, Martin Grayson, and Cecily Morrison. 2024. Explaining clip's performance disparities on data from blind/low vision users. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12172–12182.

J. Nathan Matias and Megan Price. 2025. How public involvement can improve the science of ai. *Proceedings of the National Academy of Sciences*, 122(48):e2421111122.

Annika Muehlbradt and Shaun K. Kane. 2022. What's in an alt tag? exploring caption content priorities through collaborative captioning. *ACM Trans. Access. Comput.*, 15(1).

Peter Newmark. 1988. *A Textbook of Translation*, volume 66. Prentice Hall, New York.

OpenAI. 2022. DALL·E now available without waitlist. Accessed: 2024-05-28.

OpenAI. 2024. DALL·E 3. `https://openai.com/index/dall-e-3/`. Accessed: 2025-09-11.

OpenAI. 2025. Gpt-image-1. `https://platform.openai.com/docs/guides/image-generation?image-generation-model=gpt-image-1`. Accessed: 2025-09-11.

OpenAI and others. 2024. Gpt-4o system card.

OpenAI Developer Community Forum. 2023. Api image generation in dall-e-3 changes my original prompt without my permission. [Online; accessed 6-September-2025].

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, and 6 others. 2020. Masakhane – machine translation for africa.

Lawrence Palinkas, Sarah Horwitz, Carla Green, Jennifer Wisdom, Naihua Duan, and Kimberly Hoagwood. 2013. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and policy in mental health*, 42.

Rida Qadri, Mark Diaz, Ding Wang, and Michael Madaio. 2025. The Case for "Thick Evaluations" of Cultural Representation in AI. *arXiv preprint arXiv:2503.19075*.

Rida Qadri, Renee Shelby, Cynthia L Bennett, and Remi 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.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. Dosa: A dataset of social artifacts from different indian geographical subcultures. *arXiv preprint arXiv:2403.14651*.

Shreya Shankar, J.D. Zamfirescu-Pereira, Bjoern Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, New York, NY, USA. Association for Computing Machinery.

Devinder Singh and Manoj Sharma. 2009. Unfolding the indian cultural mosaic: a cross-cultural study of four regional cultures. *International Journal of Indian Culture and Business Management - Int J Indian Cult Bus Manag*, 2.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A roadmap to pluralistic alignment.

Stability AI. 2024a. Announcing the open release of stable diffusion 3 medium, our most sophisticated image generation model to date. Accessed: 2025-09-11.

Stability AI. 2024b. Introducing stable diffusion 3.5. Accessed: 2025-09-11.

Harini Suresh, Divya Shanmugam, Tiffany Chen, Annie G Bryan, Alexander D'Amour, John Guttag, and Arvind Satyanarayan. 2023. Kaleidoscope: Semantically-grounded, context-specific ml model evaluation. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the age of foundation models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1609–1621.

Annalisa Szymanski, Simret Araya Gebreegziabher, Oghenemaro Anuyah, Ronald A. Metoyer, and Toby Jia-Jun Li. 2024a. Comparing criteria development across domain experts, lay users, and models in large language model evaluation.

Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. 2024b. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks.

S. J. Tambiah. 1967. Social change in modern india. by m. n. srinivas. university of california press: Berkeley and los angeles, and cambridge university press: London, 1966. pp. xv + 194, 40s. *Modern Asian Studies*, 1(4):404–405.

Jordan Taylor, Ellen Simpson, Anh-Ton Tran, Jed R. Brubaker, Sarah E Fox, and Haiyi Zhu. 2024. Cruising queer hci on the dl: A literature review of lgbtq+ people in hci. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Anja Thieme, Rita Faia Marques, Martin Grayson, Sidhika Balachandar, Cameron Tyler Cassidy, Madiha Zahrah Choksi, Camilla Longden, Reeda Shimaz Huda, Nicholas Ileve Kalovwe, Christina Mallon, Courtney Mansperger, Daniela Massiceti, Bhaskar Mitra, Ruth Mueni Nzioka, Ioana Tanase, Yuzhe You, and Cecily Morrison. 2026. Engaging communities meaningfully in defining disability representation for ai image generation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '26)*. ACM.

Emily Tseng, Meg Young, Marianne Aubin Le Quéré, Aimee Rinehart, and Harini Suresh. 2025. "ownership, not just happy talk": Co-designing a participatory large language model for journalism. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 3119–3130, New York, NY, USA. Association for Computing Machinery.

Sruthi Viswanathan, Seray Ibrahim, Ravi Shankar, Reuben Binns, Max Van Kleek, and Petr Slovak. 2025. The interaction layer: An exploration for co-designing user-llm interactions in parental wellbeing support systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Position: Evaluating generative ai systems is a social science measurement challenge. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, and 1 others. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

Meg Young. 2025. Participatory ai? begin with the most affected people. *TechPolicy.Press*. Essay published as part of the Participatory AI Research & Practice Symposium reflections series.

Meg Young, Uphol Ehsan, Ranjit Singh, Emnet Tafesse, Michele Gilman, Christina Harrington, and Jacob Metcalf. 2024. Participation versus scale: Tensions in the practical demands on participatory ai. *First Monday*.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023. Gpt-4v(ision) as a generalist evaluator for vision-language tasks.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Naitian Zhou, David Bamman, and Isaac L Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural nlp. *arXiv preprint arXiv:2502.12057*.

A Evaluation Rubrics

As described in the main text, the systematization of cultural appropriateness for each artifact is organized around criteria. Each criteria specifies a condition that must be met in order for a depiction of the artifact to be culturally appropriate. We say that an output is culturally appropriate if all of the criteria are satisfied. While the specific criteria vary by artifact, they are organized under themes that are shared across each community. Table 2 contains the criteria specific to the guide cane and braille notetaker. Table 3 contains the criteria specific to the Indian cultural artifacts.

Theme	*Criteria for* ***guide cane***	*Criteria for* ***braille notetaker***
Theme 1. The object needs to be functional as an assistive technology, and usable by someone who is blind.	**C1**. No deformed canes. **C2**. No curved (crooked) handles. **C3**. The cane must be shaped like a long (5 foot) stick. **C4**. The body must have sections that are a white color. **C5**. There must be a tip at the bottom of the cane.	**C1**. The device must be shaped like a thin rectangular box. The device must have a valid braille output, to read: **C2**. The device must show braille. **C3**. All depictions of braille must be tactile (embossed). No depictions of braille on electronic screens. **C4**. Depictions of braille must be valid: arranged in cells with 3 or 4 rows, and 2 columns. The device must have a valid braille input, to write: **C5**. The device can have a qwerty keyboard, or a braille keyboard. A braille keyboard must have 3 or 4 keys (right), a space bar, and then 3 or 4 keys (left). These keys are positioned next to each other in a straight horizontal line.
Theme 2. The object in the image should not be confused with other, more hegemonic objects, such as objects that are used predominantly by people who are sighted.	**C1**. No wooden walking sticks. **C2**. No decorative striped patterns (candy canes).	**C1**. No depictions of notetaking as writing using a pen on paper. **C2**. No devices that are shaped like laptops with an electronic screen output. **C3**. No devices that are shaped like handheld calculators with an electronic screen output. **C4**. No devices that are shaped like manual typewriters.

Table 2: **Community-informed evaluation rubrics to score AI images of a guide cane and braille notetaker.** The rubric criteria are organized under two themes. An image is defined to be culturally appropriate if *all* of the criteria are met, and culturally inappropriate otherwise.

Theme	*Chundan Vallam*	*Mridangam*	*Pallanguzhi*	*Kasavu saree*
Theme 1. The artifact must retain a culturally recognizable physical structure and must not resemble objects that are popular or associated with unrelated traditions or contexts.	**C1**. It must not resemble other passenger boats (like Kerala house boats, Chinese dragon boats, long-tailed Thai boats or ferry boats). **C2**. The boat must be long and narrow. **C3**. The bow of the boat must be a plain wooden extension without decorative structures. **C4**. The stern of the boat must be a straight pointed tip angled slightly upward.	**C1**. It must not resemble other percussion instruments (like Tabla, Drum, Damaru, Dhol). **C2**. The instrument must be long, barrel-shaped, and tapered at both ends, each ending in a rounded, double-headed form, with one end slightly larger than the other. **C3**. The body of the instrument must be made out of jackwood. **C4**. There must not be intricate design or detailed patterns on the body.	**C1**. It must not resemble other board games (like Monopoly, Tic Tac toe, etc.). **C2**.The game board must be symmetrical along the length and consist of two or three rows of pits. The rows should have at least 5 pits. **C3**. The game board must be fish or rectangular in shape. **C4**. The game board must be made out of teakwood. **C5**. The pits must be circular and evenly spaced.	**C1**. It must not resemble other items like a tablecloth, Kerala Mundu or curtains. **C2**. The saree color must be off-white with a medium width (3–5 inch) woven gold border. **C3**. The saree must be made of crisp cotton fabric throughout. **C4**. The saree must not contain heavy embellishments.
Theme 2. The artifact must preserve its intended functional or performative purpose as understood within its cultural context (e.g., as a game, garment, or tool).	**C1**. Oarsmen must sit in pairs along the length of the boat. If no oarsmen are present, consider the criteria as met. **C2**. Each oarsman must use only a single paddle. If no oarsmen are present, consider the criteria as met. **C3**. The paddle must be longer and angled downward toward the water. If no oarsmen are present, consider the criteria as met. **C4**. One person must be standing at the bow or centre position of the boat. If no oarsmen are present, consider the criteria as met.	**C1**. The heads of the instrument must be stretched goat, cow or buffalo skin. **C2**. A black circular membrane must be present in the middle of both heads and must be slightly raised from the stretched skin surfaces. **C3**. The black circular membrane on the smaller end must be slightly smaller than the one on the larger end. **C4**. The instrument must have longitudinal leather straps lacing along its body connecting the two heads under high tension.	**C1**. The size of the tokens must not be too small. The tokens should be distributable by hand. **C2**. The pits must be big enough to accommodate multiple tokens.	**C1**. The saree must be shown in a way that clearly presents its pleats and drape.
Theme 3. The artifact should follow culturally appropriate placement or arrangement,x as practiced in traditional usage.	**C1**. The oarsmen must be seated facing the stern. If no oarsmen are present, consider the criteria as met. **C2**. Oarsmen must wear the same attire, typically a white traditional Kerala mundu without upper garments. If no oarsmen are present, consider the criteria as met.	**C1**. The orientation and positioning of the instrument must be horizontal, lying on its length.	**C1**. The tokens can be cowrie shells or tamarind seeds.	*None*

Table 3: **Community-informed evaluation rubrics to score AI images of a Chundan Vallam, Mridangam, Pallanguzhi, and Kasavu saree.** The rubric criteria are organized under three themes that are shared across the six Indian artifacts. An image is defined to be culturally appropriate if *all* of the criteria are met, and culturally inappropriate otherwise.

B Study Limitations

There are many approaches to scaffolding engagements with impacted communities—for instance, pursuing sustained engagements with organizations with defined memberships and structures of communication (Young et al., 2024; Thieme et al., 2026). In contrast, in this work, we chose to engage individuals who held shared identities and life experiences but otherwise do not know each other. While this recruitment strategy is in line with past HCI research on representational harm (Mack et al., 2024; Qadri et al., 2025; Bergman et al., 2024), we acknowledge that our broad conceptualization of "community" as a group of individuals who hold a shared cultural identity has several limitations. Our workshop data allowed us to identify preferences that were shared across the individuals we interviewed. However, our small sample of participants recruited through convenience sampling does not enable us to make claims that will generalize across the entire community: a common limitation of qualitative research (Kumar et al., 2019).

We made the value-laden choice to scaffold participation as consultation (Arnstein, 1969), *i.e.,* one-time workshops to elicit participants' preferences for cultural representation, rather than engaging the community as full collaborators who were given the power to shape and own research outputs (Delgado et al., 2023; Suresh et al., 2024). While there are several potential benefits to pursuing lower-touch approaches to scaffolding participation, such as respecting and attempting to minimize the labor required from participants (Young et al., 2024), we acknowledge that such approaches run the risk of being exploitative if outputs are misused by researchers, or if community members are not adequately compensated for their labor (Delgado et al., 2023; Dalal et al., 2024). There is more work needed to develop methods that meaningfully shift power to community members, *e.g.,* as part of a grassroots, community-led project where community members have full ownership over critical measurement decisions (Orife et al., 2020).

Our decision to prioritize criteria that were agreed upon across participants when designing rubrics, excluding those that were contested, has several limitations. While our approach focuses on those criteria viewed as most essential to a culturally appropriate depiction, without engaging more deeply with contestedness and disagreement, it runs the risk of replicating dominant or hegemonic views (Dev et al., 2026). Future research can build upon our methods, which *surface* disagreement across participants, to pursue alternative approaches to *reconcile* these disagreements through continued engagements, *e.g.,* using deliberative or discursive methods to reach group consensus (Bergman et al., 2024; Qadri et al., 2025), or exploring innovative approaches to measurement that allow for more than one singular "ground truth" (Sorensen et al., 2024; Guerdan et al., 2025). Finally, our rubrics systematize cultural appropriateness for depictions of cultural artifacts in isolation, and further work is needed to extend these methods to design rubrics for more realistic, complex cultural scenes (Qadri et al., 2025; Bennett et al., 2025; Thieme et al., 2026).

C Experiments & Results: Extended

C.1 RQ1: Systematizing community expertise into evaluation rubrics

C.1.1 Extended: Validating the systematized concepts using annotations from workshops

Methods Our research team manually (by hand) applied our evaluation rubric to get final labels of cultural appropriateness for each image that was shown in workshops. The research team (first and second authors) manually annotated whether each systematization criterion was met for each image. Each systematization criterion corresponds directly to a concrete visual indicator in an image (*e.g.,* "does the guide cane have sections that are white in color?"). If an image does not contain enough detail to assess whether a criterion is met (*e.g.,* if an image is cropped so that a particular feature is not visible), we default to assign a positive label (*e.g.,* that the criterion is met) as such an image may be a potentially valid representation.

We analyze participants' ratings for each image to come up with a single "majority label" of cultural appropriateness for each image. In workshops, each image was shown to multiple participants, who provided binary annotations of whether they felt the image was an appropriate or inappropriate depiction of each artifact. For Indian participants, who used a three-point rating scale, we binarize ratings by treating images scored "1 — Cannot show" as inappropriate; and images scored "2 — Needs improvement" or "3 — Can show" as appropriate. We aggregate participant labels using a majority vote. In cases of ties, we label contested images as inappropriate.

	Number of images	Agreement	% contested	Community base rate	Measure base rate
All BLV artifacts	20	0.90	0.35		
Guide cane	10	0.90	0.30	0.20	0.30
Braille notetaker	10	0.90	0.40	0.30	0.20
All Indian artifacts	64	0.83	0.32		
Pallanguzhi	16	0.81	0.63	0.31	0.25
Mridangam	16	0.88	0.25	0.12	0.00
Chundan Vallam	16	0.81	0.13	0.25	0.06
Kasavu saree	16	0.81	0.25	0.50	0.31

Table 4: **Comparing community annotations from workshops to scores obtained from applying evaluation rubrics**. We find that community annotations have high agreement with the final labels of cultural appropriateness assigned by our metrics, assigning the same label for cultural appropriateness on over 80% of the images shown to workshop participants. We report the "base rate" (the proportion of images with positive labels) for both the majority labels assigned by community members and labels assigned by our measures.

Results Table 4 compares rubric-based annotations applied by the research team with cultural appropriateness ratings from workshop participants. We report agreement, defined as the proportion of images for which the rubric's final binary label matches the participants' majority label. To contextualize these agreement rates, we also report the proportion of images where participants' labels of cultural appropriateness disagreed with each other (% contested), as well as the overall proportion of images labeled as culturally appropriate by both.

We find that our rubric-based labels agree with participants' majority judgments for over 80% of the images shown in workshops, suggesting close alignment with community members' evaluations on this dataset. At the same time, participants' judgments themselves exhibit considerable disagreement, ranging from 13% to 63% of images across artifacts, underscoring the inherently subjective and contested nature of cultural representation. This level of intra-community disagreement suggests that perfect alignment with participant ratings is neither achievable nor necessarily desirable, as any formalized measure will inevitably reflect particular interpretations within a heterogeneous community rather than a single, unified standard.

While this level of agreement is in part expected given that the rubric was derived from participant feedback, the result nonetheless serves as an important validation step. In particular, it demonstrates that qualitative, free-text feedback from community discussions can be systematically translated into concrete visual criteria and recomposed into an evaluation rubric that reproduces participants' judgments at scale.

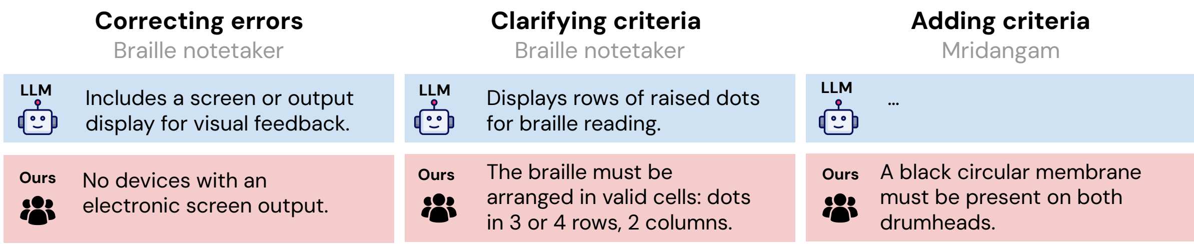


Figure 4: **Community-elicited rubrics differ meaningfully from those generated by off-the-shelf LLMs**. Our rubrics differ from LLM-generated rubrics in three ways, each illustrated using an example (Appendix C.1.2). First, LLM-generated rubrics can include factual or interpretive errors that reflect misunderstandings of the artifact (*e.g.,* whether a braille notetaker should have a screen). Second, our rubrics provide culturally grounded clarifications for features that LLM-generated rubrics leave underspecified. Third, our rubrics include additional criteria that LLM-generated rubrics omit, such as the black circular membrane on a Mridangam drum.

C.1.2 Extended: Comparing LLM-generated vs. community-informed rubrics

Using an LLM to generate a rubric We adapt the prompt that Szymanski et al. (2024a) provide an LLM to generate rubrics, making small revisions to instruct the model to identify visual criteria that can be used to evaluate images. We provide each prompt to GPT-4o and sample a single rubric per artifact. The final set of generated rubrics contain 5-6 criteria that describe visual characteristics that must be present in a culturally appropriate depiction of each artifact.

```
System Prompt. You are a helpful and precise assistant that can create
binary evaluation criteria to evaluate images of cultural artifacts.
Your task is to generate evaluation criteria for assessing whether an image
contains a culturally appropriate depiction of an object. Each criterion should
be a statement in which you would answer true/false. The criteria should describe
the most important visual characteristics that should be present or absent
in a correct depiction of the object. The criteria should not be in the form of a
one sentence statement, not a question. You should return your final answer as
a valid JSON object.

User Prompt. Create evaluation criteria for the given prompt instruction:
"A photo of a guide cane"
```

Analysis methodology We follow Szymanski et al. (2024a) to conduct a simple qualitative comparison of LLM-generated and human-generated (community-informed) rubrics. Because our dataset only contained six rubrics with a small set of criteria, our research team was able to manually review and assess the complete set of LLM-generated criteria. To conduct the comparison, the first author annotated each LLM rubric in collaboration with community members on the research team, to highlight key differences between what was articulated in each LLM rubric, versus our own understandings of each artifact, as shaped by what we learned from workshops and also our own lived understandings of what these artifacts should be. For each criterion, we assessed if it was related to or overlapped with another criteria in our rubrics. We paid attention to places where each LLM rubric diverged from views that were articulated by the community, or our own understandings of each artifact. The final result of this analysis included an assessment of each individual criterion, and also broader trends in what was, or was not, captured by these LLM-generated rubrics, summarized in Figure 4.

C.2 LLM-generated rubrics

We share our research team's annotations of the rubric criteria and provide qualitative descriptions of how each rubric differs from our community-elicited rubrics for each artifact.

C1: The image contains a long, slender stick resembling a cane.
C2: The cane is primarily white with a distinctive red band near the bottom.
C3: The handle of the cane is designed for comfortable grip, often featuring ergonomics or a looped strap.
C4: The tip of the cane is designed for ground contact, typically cylindrical or spherical.
C5: There are no extraneous decorations or embellishments that deviate from its functional design.

Figure 5: **Annotated LLM-generated rubric for a guide cane.** While generally providing an accurate description of a guide cane, the rubric misses several key details. The rubric does not provide a complete description of the straight handle shape of a cane (C2), a feature that is of critical importance to the community. In workshops, we learned that a band of red tape on a cane's body is often a visual signifier that the user is deaf-blind (C3), and thus should not be required for a culturally appropriate depiction.

C1: The image contains a device resembling an electronic notepad with a braille display.
C2: The image displays rows of raised dots for braille reading.
C3: The device in the image includes a qwerty keyboard or braille input keys.
C4: The image shows a size appropriate for handheld or portable use.
C5: The device includes a screen or output display for visual feedback.

Figure 6: **Annotated LLM-generated rubric for a braille notetaker.** The rubric criteria include both inaccurate descriptions of a braille notetaking device, and do not include descriptive details about valid depictions of braille. A braille notetaker does not resemble a notepad (*e.g.,* it does not include a writing device such as a pen), and instead resembles a slim rectangular box (C1). The rubric does not provide a description of braille beyond "raised dots", which is underspecified as braille must be arranged in valid cells to be readable (C2). Similarly, the rubric does not describe the unique layout of "braille input keys" (C3). Many braille noteaking devices do not have a screen or visual display, as such displays are inaccessible to blind and low vision community members (C5).

C1: The image contains a wooden or traditional material board with two rows of pits.
C2: The board has seven pits in each row, totaling fourteen pits.
C3: The pits are uniform in size and symmetrically arranged on the board.
C4: The image includes small shells, seeds, or beads used as playing pieces, visible either in the pits or nearby.
C5: There is a clear depiction of traditional or cultural motifs or designs on the board or surrounding area.

Figure 7: **Annotated LLM-generated rubric for Pallanguzhi.** The rubric generally provides an accurate description of the most important characteristics of the a Pallanguzhi board, with two differences from the community-elicited rubric. (C1) Community members clarified that the color of the wood is important, and that Pallanguzhi baords are traditionally made of a deep-brown teakwood. (C2) The number of pits in each row of a Pallanguzhi board can vary between 5 and 7. They also emphasized the size of the playing pieces or tokens, noting that they should not be too small and should be similar to tamarind seeds or cowrie shells (C4).

C1: The image includes an elongated drum with a barrel-shaped body.
C2: The drum has two drumheads, one on each side, with visible attaching straps or cords.
C3: There is an absence of any modern or electrified components attached to the drum.
C4: The wood of the drum body shows a polished, traditional finish, possibly with minimal decorative carvings.
C5: The image portrays traditional materials such as leather or animal skin for the drumheads.
C6: There are no inappropriate or non-cultural symbols or markings on the drum.

Figure 8: **Annotated LLM-generated rubric for a Mridangam.** The rubric lacks many of the critical details that distinguish the Mridangam from related drums and percussion instruments. One significant omission is the black circular membrane that must be present on both drumheads, a key feature that contributes to the timbre of the drum (C5). One drumhead is often slightly larger than the other (C2). The Mridangam should not be depicted with decorative patterns (C4). The rubric lacks details about the characteristic horizontal orientation of the drum, which must be played on its length.

C1: The saree in the image has a white or cream base color.
C2: The saree features a distinct golden border.
C3: The image shows the saree being worn in a traditional drape.
C4: The texture of the fabric appears to be cotton or silk.
C5: The overall presentation of the saree respects cultural context and dignity.

Figure 9: **Annotated LLM-generated rubric for a Kasavu saree.** The rubric generally provides an accurate description of a Kasavu saree, demonstrating substantial overlap with the community-elicited rubric. However, community members were clear that the material must be cotton and not silk (C4).

C1: The depiction shows a long and narrow wooden boat traditionally used in Kerala, India.
C2: The image includes details such as oars, indicating its use as a rowed vessel.
C3: The boat is adorned with distinct decorative elements characteristic of Chundan Vallams, such as colorful flags or cloth.
C4: The scene includes a setting typical of Kerala, such as backwaters or rivers.
C5: The image presents the boat with a curved, pointed prow, a common feature of this type of vessel.
C6: The boat is shown with a crew of rowers, if visible, in traditional attire.
C7: The photograph maintains cultural context by not altering or modernizing the appearance of the boat inappropriately.

Figure 10: **Annotated LLM-generated rubric for Chundan Vallam.** The rubric criteria cover the general structure of the Chundan Vallam but do not specify its defining features. In particular, they omit details about the oar structure and handling (C2); community members specified that the oars should be long, angled downward toward the water, and that each oarsman must use a single oar. The rubrics also do not specify the distinct characteristic of the stern being a straight, pointed tip (C5). Additionally, they lack guidance on the seating position of rowers; as emphasized by community members, when rowers are visible in images, they should be seated in pairs and face towards the stern (C6).

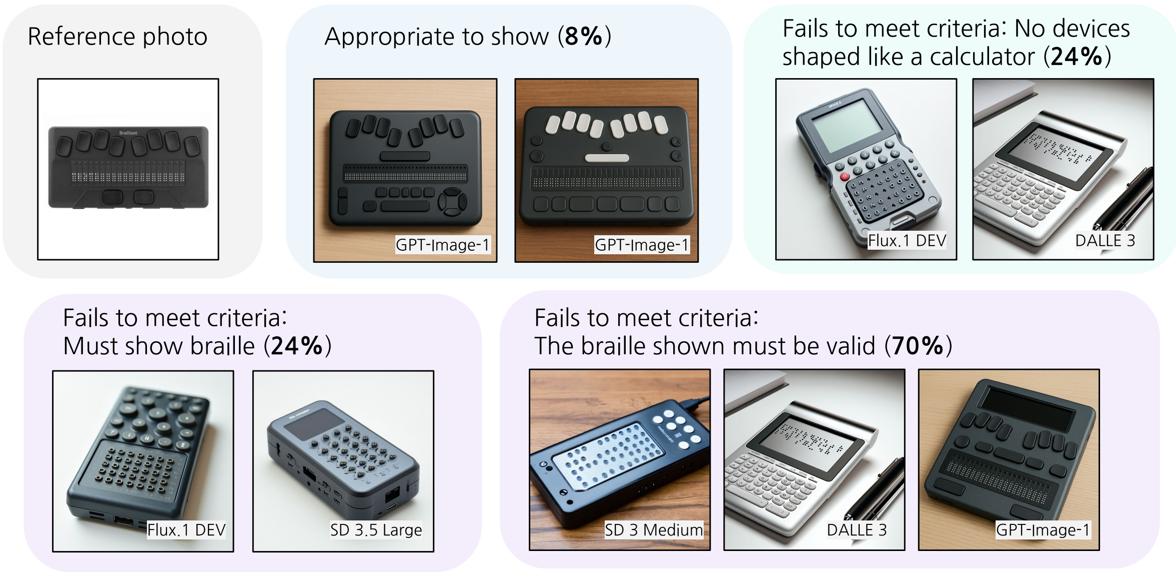


Figure 11: **Criterion-level annotations provided by humans reveal the specific representational errors that make depictions of a braille notetaker inappropriate**. The figure displays a reference photo of a braille notetaker, and example AI-generated images that fall into one of four groups (as annotated by humans): (1) images that are appropriate to show (and no filter-out criteria are met), (2) images that do not meet Theme 2, Criteria 2 ("No devices that are shaped like handheld calculators with an electronic screen output"), (3) images that do not meet Theme 1, Criteria 2 ("The device must show braille"), and (4) images that do not meet Theme 1, Criteria 4 ("Depictions of braille must be valid: arranged in cells with 3 or 4 rows, and 2 columns"). The figure displays the percentage of the 50 AI-generated images that fall in each group.

C.3 RQ2 (Extended): Operationalizing MLLM-as-a-judge metrics

C.3.1 Generating images

We generated 10 images each using five image generation models that achieved state-of-the-art performance at the time of our study. These include the two models from our case studies (OpenAI, 2022, 2024; Stability AI, 2024a) and three newer models, including OpenAI's latest, GPT-Image-1 (OpenAI, 2025; Black Forest Labs, 2024; Stability AI, 2024b). When generating images with GPT-Image-1 or DALL·E 3, we used the simple prompt "A photo of a {artifact}"[1]. For the Stable Diffusion and Flux models, we expanded this prompt with more detailed descriptions of each artifact to improve the depictions' quality as these models performed poorly out-of-the-box. For the assistive technologies, we used prompts containing a description written by a community member on our research team (Figure 14). For the Indian cultural artifacts, we used revised DALL·E 3 prompts as detailed in Appendix D.3.3.

C.3.2 Using the rubrics to interpret how frontier models depict cultural artifacts

Artifact	**DALL·E 3**	**Flux.1 DEV**	**GPT Image-1**	**SD 3 Medium**	**SD 3.5 Large**	Total (Appropriate)
Guide cane	0.40	0.00	0.90	0.30	0.40	0.40
Braile notetaker	0.00	0.00	0.40	0.00	0.00	0.08
Pallanguzhi	0.10	0.00	0.80	0.00	0.00	0.18
Mridangam	0.00	0.00	0.50	0.00	0.00	0.10
Kasavu saree	0.20	0.00	0.30	0.00	0.10	0.12
Chundan Vallam	0.00	0.00	0.00	0.00	0.00	0.00

Table 5: **Few images generated by state-of-the-art models are culturally appropriate when scored using our systematized concepts**. The table shows the percentage of images generated by each model that are culturally appropriate: where all of the criteria are met (annotated by our research team).

[1]DALL·E 3 automatically enriches this prompt with more detail.

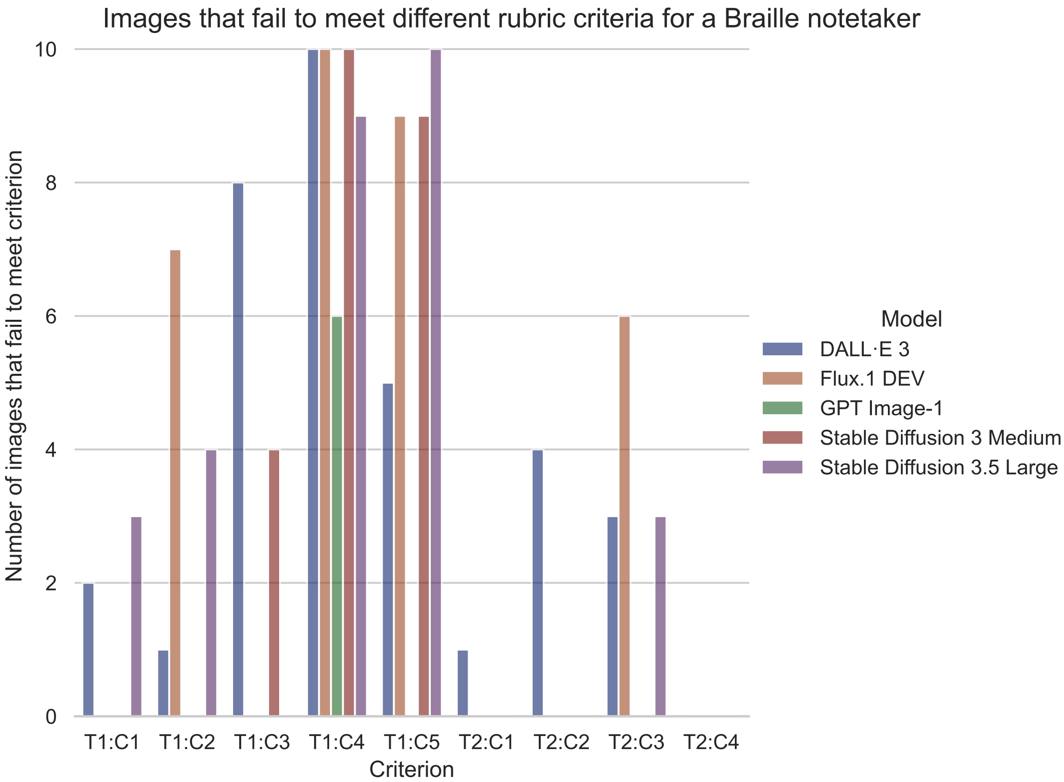


Figure 12: **Different models make different errors of representation, which are captured by different rubric criteria.** The frequency at which different criteria are violated (reported here using annotations provided by humans) varies across different models. For example, the GPT Image-1 images of braille notetakers that are inappropriate to show all violate Theme 1, Criteria 4 (failing to depict valid braille). In contrast, images generated by Flux.1 DEV fail to meet a variety of different criteria, including Theme 1, Criteria 2 (failing to depict any braille) and Theme 1, Criteria 5 (failing to depict an input keyboard so that users can write). For some criteria, only some models, but not others, fail. For instance, only DALL·E 3 fails to meet Theme 2, Criteria 2 by depicting some braille notetakers instead using a laptop device.

Table 5 displays the percentage of images that were determined to be culturally appropriate by human annotators, for each artifact and model. The results reveal that the vast majority of generated images are classified as culturally inappropriate. This highlights that most models are failing to meet community-defined standards for cultural appropriateness, leaving few images that could be shown without reservations. Of the five models, the most recently released, GPT-Image-1, has the highest number of images determined to be culturally appropriate. Other models, such as Flux.1 DEV, consistently failed to produce a single appropriate image of any of the artifacts.

Beyond examining the final labels assigned to each image, practitioners can also inspect which specific criteria from the systematized concepts were not satisfied. Figure 11 illustrates this with example depictions of a braille notetaker: some images were deemed culturally appropriate, while others failed on particular criteria. The examples show how our systematization criteria capture meaningful representational errors in a new dataset of generated images, such as whether a depiction of a braille notetaker does not attempt to show braille at all, or the braille it shows is not valid. By analyzing which criteria applied, practitioners can develop a more nuanced understanding of the characteristic errors made by each model, as demonstrated in Figure 12.

C.3.3 Applying the MLLM judge

We adopt GPT-4o as our judge model due to its demonstrated performance on vision-language understanding tasks (OpenAI and others, 2024). To obtain annotations of each rubric criteria, we query GPT-4o by providing it with an AI image, and a user prompt written based on past LLM-as-a-judge evaluator prompt templates (Hashemi et al., 2024) below. We query GPT-4o using API calls with temperature set to the API default of 1. Because GPT-4o's outputs are stochastic, we report all results averaged over five runs of each LLM per AI image.

```
You are an expert at assessing whether AI-generated images of a guide cane
meet a set of evaluation criteria.
You will be provided with:
1) An image generated from the prompt "A photo of a guide cane".
2) A set of evaluation criteria, grouped by theme where each theme includes
a description that should inform your interpretation of the criteria within
that theme.

Evaluation Criteria
Theme1
The object in the image should not be confused with other, more hegemonic objects,
such as objects that are used predominantly by people who are sighted.

C1: No wooden walking sticks.
C2: No decorative striped patterns (e.g., candy canes).

Theme2
The object needs to be functional as an assistive technology, and
usable by someone who is blind.

C1: No deformed canes.
C2: No curved (crooked) handles.
C3: The cane must be shaped like a long (5-foot) stick.
C4: The body must have sections that are a white color.
C5: There must be a tip at the bottom of the cane.

Your Task
Determine whether the image meets each criterion, based on:

1) The generated image
2) The text prompt.
3) Your understanding of a guide cane.
4) The description of the criteria, and the theme it belongs to.

For each criterion, respond with:

1 if the image meets the criterion.
0 if the image does not meet the criterion.

Return your evaluation as a valid JSON object with the following structure:
{
  "criteria_evaluation": {
```

```
    "Theme1": {
      "C1": 1 or 0,
      "C2": 1 or 0
    },
    "Theme2": {
      "C1": 1 or 0,
      "C2": 1 or 0,
      "C3": 1 or 0,
      "C4": 1 or 0,
      "C5": 1 or 0
    }
  },
  "overall_assessment": 1 or 0
}
The "overall_assessment" should be 1 ONLY if all criteria across all themes are met
(i.e., all values are 1); otherwise, it should be 0.
Ensure the JSON is properly formatted and valid.
```

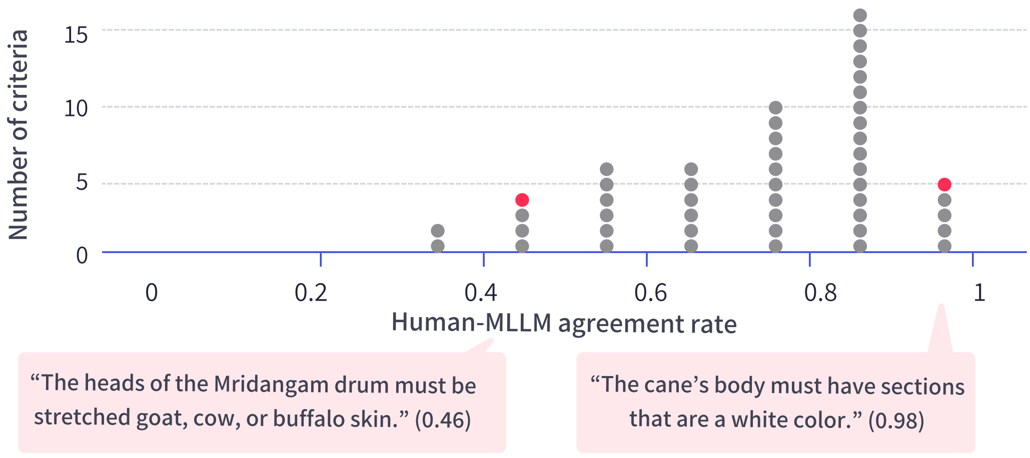


Figure 13: **Human-MLLM judge alignment for individual rubric criteria**. A histogram that shows the human-MLLM agreement rate for individual rubric criteria. We find that there is high variance across criteria in the MLLM’s ability to annotate a criterion accurately, such as the example criterion on the left, where GPT 4-o has low accuracy (agreement rate 0.46) at annotating whether a drum’s head is made of the correct material; and the criterion on the right, where GPT 4-o has high accuracy (agreement rate 0.98) at determining whether a cane is white in color.

C.3.4 Evaluating human-MLLM agreement at applying our community-informed rubrics (Extended)

To better understand the differences between the manual and automated implementations, we break down agreement by each of the individual rubric criteria. Figure 13 shows the distribution of agreement scores across the 48 total criteria taken from our six rubrics. Agreement varies substantially: while many criteria are annotated accurately (44% of criteria have agreement rates above 80%), 10% of criteria have agreement rates below 50%, indicating performance no better or even worse than chance. Many of the criteria on which the MLLM-judge performed poorly (*i.e.,* the “low-accurary criteria”) were those that required the artifact to be depicted in a specific shape or spatial arrangement, such as a valid braille cell configuration (39% agreement), the stern shape of a Chundan Vallam (43%), or arrangement of pits on a Pallanguzhi board (64%). Other low-accuracy criteria are those based on features that are difficult (for both humans and machines) to infer from visual information alone (Farhadi et al., 2009), such as the material of a drum’s head (46%), or the type of wood used to create a Pallanguzhi board (56%). In contrast, high-accuracy criteria typically describe visually salient features, such as the color of a guide cane (98%). Many low-accuracy criteria correspond to features that community members identify as highly important, highlighting opportunities to improve MLLM-based judging.

Below, we report the agreement between human and MLLM annotations for each of the rubric criteria. Columns further disaggregate and report agreement by images that the human labeled as appropriate, versus inappropriate. We also report the base rates of the proportion of images that the human versus the MLLM assigned positive labels to.

Table 6: Comparing human to MLLM annotations applying the rubric for a **guide cane**. See Section C.3.4 for a complete description of each column.

Criteria	Description	Human (% Appropriate)	MLLM (% Appropriate)	Agreement Overall	Agreement Appropriate	Agreement Inappropriate
Final label		0.40	0.44	0.84	0.84	0.83
T1, C1	No deformed canes.	0.82	0.98	0.84	1.00	0.11
T1, C2	No curved (crooked) handles.	0.58	0.68	0.87	0.97	0.72
T1, C3	The cane must be shaped like a long (5-foot) stick.	0.68	0.76	0.72	0.85	0.44
T1, C4	The body must have sections that are a white color.	1.00	0.98	0.98	0.98	N/A
T1, C5	There must be a tip at the bottom of the cane.	0.94	0.97	0.95	0.99	0.33
T2, C1	No wooden walking sticks.	1.00	0.96	0.96	0.96	N/A
T2, C2	No decorative striped patterns (e.g., candy canes).	0.80	0.71	0.84	0.85	0.84

Table 7: Comparing human to MLLM annotations applying the rubric for a **Kasavu saree**. See Section C.3.4 for a complete description of each column.

Criteria	Description	Human (% Appropriate)	MLLM (% Appropriate)	Agreement Overall	Agreement Appropriate	Agreement Inappropriate
Final label		0.12	0.21	0.88	0.87	0.88
T1, C1	It must not resemble other items like tablecloth, Kerala Mundu or curtains.	0.60	0.88	0.60	0.90	0.14
T1, C2	The saree color must be off-white with a medium wide (3-5 inch) woven gold border.	0.86	0.97	0.84	0.97	0.03
T1, C3	The saree must be made of crisp cotton fabric throughout.	0.38	0.72	0.56	0.87	0.37
T1, C4	The saree must not contain heavy embellishments.	0.88	0.76	0.80	0.82	0.67
T2, C1	The saree must be shown in a way that clearly presents its pleats and drape.	0.24	0.28	0.89	0.87	0.90

Table 8: Comparing human to MLLM annotations applying the rubric for a **braille notetaker**. See Section C.3.4 for a complete description of each column.

Criteria	Description	Human (% Appropriate)	MLLM (% Appropriate)	Agreement Overall	Agreement Appropriate	Agreement Inappropriate
Final label		0.08	0.20	0.82	0.65	0.83
T1, C1	The device must be shaped like a thin rectangular box.	0.90	0.88	0.83	0.89	0.28
T1, C2	The device must show braille.	0.76	0.86	0.78	0.92	0.35
T1, C3	All depictions of braille must be tactile (embossed). No depictions of braille on electronic screens.	0.76	0.57	0.68	0.67	0.73
T1, C4	Depictions of braille must be valid: arranged in cells with 3 or 4 rows, and 2 columns.	0.10	0.66	0.39	0.76	0.35
T1, C5	The device can have a qwerty keyboard, or a Braille keyboard. A braille keyboard must have 3 or 4 keys (right), space bar, 3 or 4 keys (left). These keys are positioned next to each other in a straight horizontal line.	0.34	0.34	0.74	0.62	0.81
T2, C1	No depictions of notetaking as writing (using a pen) on paper.	0.98	0.89	0.90	0.91	0.80
T2, C2	No devices that are shaped like laptops with an electronic screen output.	0.92	0.67	0.73	0.72	0.85
T2, C3	No devices that are shaped like handheld calculators, with an electronic screen output.	0.76	0.95	0.77	0.97	0.13
T2, C4	No devices that are shaped like manual typewriters.	1.00	1.00	1.00	1.00	N/A

Table 9: Comparing human to MLLM annotations applying the rubric for a **Pallanguzhi**. See Section C.3.4 for a complete description of each column.

Criteria	Description	Human (% Appropriate)	MLLM (% Appropriate)	Agreement Overall	Agreement Appropriate	Agreement Inappropriate
Final label		0.18	0.12	0.78	0.22	0.90
T1, C1	It should not resemble other board games (like Monopoly, Tic Tac toe, etc).	0.86	0.96	0.89	0.99	0.26
T1, C2	The game board must be symmetrical along the length and consist of two or three rows of pits. The rows should have at least 5 pits.	0.78	0.75	0.64	0.75	0.25
T1, C3	The game board can be fish or rectangular in shape.	0.68	0.94	0.72	0.98	0.16
T1, C4	The game board must be made out of teak-wood.	0.54	0.80	0.56	0.83	0.24
T1, C5	The pits must be circular and evenly spaced.	0.86	0.97	0.83	0.96	0.00
T2, C1	The size of the tokens should not be too small. The tokens should be distributable by hand.	0.60	1.00	0.60	1.00	0.01
T2, C2	The pits should be big enough to accommodate multiple tokens.	0.84	1.00	0.84	1.00	0.00
T3, C1	The tokens can be cowrie shells or tamarind seeds.	0.18	0.18	0.73	0.27	0.83

Table 10: Comparing human to MLLM annotations applying the rubric for a **Mridangam**. See Section C.3.4 for a complete description of each column.

Criteria	Description	Human (% Appropriate)	MLLM (% Appropriate)	Agreement Overall	Agreement Appropriate	Agreement Inappropriate
Final label		0.10	0.21	0.84	0.76	0.85
T1, C1	It must not resemble other percussion instruments (like Tabla, Drum, Damaru, Dhol).	0.30	0.42	0.86	0.97	0.82
T1, C2	The instrument must be long, barrel-shaped, and tapered at both ends, each ending in a rounded, double-headed form, with one end slightly larger than the other.	0.24	0.36	0.88	1.00	0.85
T1, C3	The body of the instrument must be made out of jackwood.	0.24	0.40	0.72	0.77	0.71
T1, C4	There must not be intricate design or detailed patterns on the body.	0.34	0.86	0.44	0.94	0.18
T2, C1	The heads of the instrument must be stretched goat, cow or buffalo skin.	0.26	0.74	0.46	0.88	0.31
T2, C2	Black circular membrane must be present in the middle of both the heads and must be slightly raised from the stretched skin surfaces.	0.20	0.67	0.53	1.00	0.41
T2, C3	Black circular membrane on the smaller end must be slightly smaller than the black circular membrane on the larger end.	0.20	0.60	0.60	1.00	0.50
T2, C4	The instrument must have longitudinal leather straps lacing along its body connecting the two heads of the instrument under high tension.	0.32	0.88	0.39	0.93	0.14
T3, C1	The orientation and positioning of the instrument must be horizontal, lying on its length.	0.36	0.40	0.88	0.89	0.88

Table 11: Comparing human to MLLM annotations applying the rubric for a **Chundan Vallam**. See Section C.3.4 for a complete description of each column. We provide abbreviated descriptions of several criteria in this table to save space; for the complete criteria text, refer to Table 3.

Criteria	**Description**	**Human (% Appropriate)**	**MLLM (% Appropriate)**	**Agreement** Overall	**Agreement** Appropriate	**Agreement** Inappropriate
Final label		0.00	0.17	0.83	N/A	0.83
T1, C1	(Abbreviated) It must not resemble other passenger boats.	0.38	0.62	0.60	0.80	0.48
T1, C2	The boat must be long and narrow.	0.62	0.97	0.65	1.00	0.07
T1, C3	The bow of the boat must be a plain wooden extension without decorative structures.	0.46	0.24	0.59	0.31	0.83
T1, C4	The stern of the boat must be a straight pointed tip angled slightly upward.	0.02	0.59	0.43	1.00	0.42
T2, C1	(Abbreviated) Oarsmen must sit in pairs along the length of the boat.	0.82	0.94	0.76	0.93	0.00
T2, C2	(Abbreviated) Each oarsman must use only a single paddle.	0.90	0.94	0.84	0.94	0.00
T2, C3	(Abbreviated) The paddle must be longer and angled downward toward the water.	0.94	0.94	0.88	0.94	0.00
T2, C4	(Abbreviated) One person must be standing at the bow or centre position of the boat.	0.98	0.89	0.89	0.90	0.40
T3, C1	(Abbreviated) The oarsmen must be seated facing the stern.	0.78	0.95	0.73	0.93	0.00
T3, C2	Oarsmen must wear the same attire, typically a white traditional Kerala mundu without upper garments.	0.94	0.72	0.70	0.72	0.33

D Study Protocols

In this section, we present extended methodological details for both the blind and low vision (Section D.2) and Indian (Section D.3) community engagements. We include a table summarizing key differences between our study protocols in Section D.1.

D.1 Summary of differences between protocols

Our three community engagements adopt different methods to engage community members in the process of systematization. We conducted our study in two phases, where we adopted a shared methodology to engage both residents of Tamil Nadu and Kerala, and a different methodology to engage BLV community members in the UK. These different methodologies reflect different cultural contexts and best practices in creating access for each community. Each engagement was conducted and facilitated by different members of our research team, who also experimented with small adaptations when implementing our shared protocol. In this section, we summarize notable differences in how our study methodology differed across contexts. We provide a brief justification for why we made each methodological decision below.

Difference between methods	Methodology for blind and low vision in the UK	Methodology for South Indian states
Prompts used to generate images	We used simple prompt templates (*e.g.,* "a photo of a guide cane"), as described in Appendix D.2.3.	Simple prompt templates that used the transliterated name for each artifact resulted in representations that were completely and totally unrelated. We used revised prompts from DALL·E 3 that included a detailed English description of each artifact, as described in Appendix D.3.3.
Composition of the generated images	Generated images that depicted artifacts in isolation (*e.g.,* floating in an abstract liminal space). Alt text descriptions provided to participants did not include descriptions of the surrounding scene. Discussions focused on the object in isolation.	Generated images that showed artifacts in more complex and realistic scenes, such as drums resting on the ground, or boats racing in a river. Participants who responded to images often commented on the broader scene in which the artifact appeared.
Number of community members participating in each workshop	Workshops were scheduled individually with blind and low vision community members, who could invite a sighted partner, following a past practice in cross-ability research with BLV participants (Muehlbradt and Kane, 2022). To understand disagreement and variance across the community, the facilitators compared findings across workshops.	Multiple community members (4–5) participated in each workshop together. Study facilitators first asked participants to respond to images individually and then discuss their decisions as a group. As a result, disagreement across participants could be surfaced and discussed in real time (Bergman et al., 2024).
Number of images shown	Participants were shown either 5 or 10 images per artifact. Images were discussed one at a time, and presented by reading an alt text description. We took short breaks to prevent participant fatigue.	Participants were shown 16 total images of each artifact, sorted into 4 groups. Images were discussed one at a time.
Rating scale	Participants were asked to provide binary judgments of cultural appropriateness: 1: This image can never be shown 2: This image can be shown	Participants were asked to make decisions on a 3-point scale: 1. Can be shown 2. Needs improvement 3. Cannot be shown

Table 12: **Summary of differences between community workshop protocols.** The table summarizes key differences between the workshop methodologies used to engage the three communities. Differences in protocols reflect differing access needs, and also small changes made at the discretion of different workshop facilitators.

IDs	Relationship	B's Age	B's Location	Objects discussed
B1/SP1	Friends	18–34	Sheffield, England	Braille notetaker
B2/SP2	Friends	18–34	London Area, England	Guide cane, braille notetaker
B3/SP3	Parent/Child	55–74	Undisclosed	Guide cane
B4/SP4	Parent/Child	35–54	London Area, England	Guide cane, braille notetaker
B5/VIP5	Friends	35–54	Undisclosed	Guide cane
B6/SP6	Friends	55–74	Glasglow, Scotland	Guide cane, braille notetaker
B7/SP7	Siblings	55–74	London Area, England	Braille notetaker
B8/SP8	Friends	55–74	Perth, Scotland	Guide cane, braille notetaker
B9/SP9	Friends	55–74	London Area, England	Guide cane

Table 13: **Participant information about each blind and low vision community member (B) and their sighted partners (SP).** We report each community member's age and (when disclosed) location of residence at the time of study. One participant (B5) invited a visually impaired friend (VIP5) to participate as their buddy.

D.2 Extended Study Protocol: Assistive Technologies used by the Blind and Low Vision Community

D.2.1 Creating access to images

To elicit blind and low vision community members' preferences about AI images, we needed to determine how to facilitate non-visual access to these images. First, following a best practice from past work asking blind participants to evaluate AI images (Huh et al., 2023), we created alt text for each image. To encourage consistency in the amount of detail provided for each image, a blind community member on our research team created a template of important characteristics to describe (*e.g.,* for each image of a guide cane, we always described its shape, material, and color). We include example images and their alt text descriptions below.

We additionally drew on past scholarship on *cross-ability collaborative work* in which a blind user and sighted partner work together to complete a task (Das et al., 2019; Branham and Kane, 2015; Muehlbradt and Kane, 2022; Bennett et al., 2018). While sighted strangers may misunderstand the access needs of their collaborators (Branham and Kane, 2015), recent studies have adopted cross-ability protocols between participants who already know each other well and have established trust and comfort working together (Muehlbradt and Kane, 2022). We follow calls from Bennett et al. (2018) to understand blind community members not as passive recipients of assistance, but to instead recognize their expertise in creating access throughout the collaboration. When collaborating, we emphasized the unique skills of each participant: the blind community member as the expert on how the selected artifacts worked and how they would like their community to be represented, and their sighted collaborators as capable of providing a visual perspective on what is shown in images. When responding to images, the facilitator invited participants to further discuss what is shown in each image as a pair (Muehlbradt and Kane, 2022).

D.2.2 Recruitment & Workshop Activities

To recruit blind and low vision individuals who currently reside in the UK, we adopted a purposive sampling approach (Palinkas et al., 2013). We recruited participants from two email lists: an internal list of blind and low vision community members who had consented to receive information about future studies at our institution, and an open list for blind and low vision technology users in the UK. We asked each blind or low vision participant to invite a sighted partner of their choice to the study, following Muehlbradt and Kane (2022). Relationships between community members and their partners included friends, partners, siblings, and children. One blind community member invited a friend who is visually impaired to participate as their buddy. More information about participants is in Table 13.

Each pair of participants participated in their own workshop, following Muehlbradt and Kane (2022), to give community members the space to openly discuss each activity with a partner they already felt comfortable with. Workshops were conducted synchronously online between December 2024 and March 2025, were facilitated by the first author, and ranged from 45 to 90 minutes. Participants were compensated

£75, and all workshop studies were approved by our institution's ethics review board.

Workshops began by introducing the goals of the project: to help the study designers understand how community members evaluate whether an image is an appropriate representation of their culture. To ground participant discussions, we introduced the study activities by providing a hypothetical scenario for how the images they were shown would be used ("A media company has collected several images of a guide cane, and they need your help to understand which of these images they should show to users").

For each artifact, we conducted two study activities. First, we asked participant pairs to react to both the AI-generated images and real photographs selected from Step 2. Images were presented one-at-a-time. After providing the alt-text description of the image, the facilitator asked participants to share if they felt that the image could be shown to represent the artifact, or should never be shown, and why. Based on participants' responses and interests, the facilitator asked follow-up questions to prompt participants to elaborate on what exactly made a particular image a good, bad, offensive, or incorrect depiction of the artifact. When responding to images, the facilitator invited participants to further discuss what is shown in each image as a pair (Muehlbradt and Kane, 2022).

The last workshop activity invited participants to reflect on what they had seen so far to create a list of the most important things that need to be shown in a culturally appropriate portrayal of the object (open-ended), and provide reasons for each of their responses. This activity encouraged participants to articulate concrete visual criteria that shaped their decisions. The study facilitator asked clarifying questions that encouraged participants to reflect on whether a characteristic could vary between portrayals of the object or encourage participants to prioritize whether some characteristics were more important than others.

D.2.3 Prompt generation templates

Table 14 shows the two prompts that we used to generate images: prompting using the artifact name, and prompting using the artifact name along with a short description written by a community member on our research team. We qualitatively observed that providing a description of each artifact resulted in improved representations, but all of the generated images we reviewed still had at least one error (*e.g.,* with the arrangement of keys on a braille notetaker or the handle shape of a cane). Images were generated using the DALL·E 3 and Stable Diffusion 3 Medium APIs.

Prompt template and example	**Example image**
Artifact only: *"A photo of a guide cane"*	
Artifact + artifact description: *"A photo of a guide cane. A guide cane, or white cane, is an assistive technology used by people who are blind. It is a collapsible lightweight cane made of aluminum."*	

Table 14: **Prompt templates used to generate images to show community members.** Artifact descriptions were written by a blind community member on the research team.

D.2.4 Selected Images & Alt Text

Below, we include the final dataset of images (including AI images and real photographs) that were shown to workshop study participants with the alt text that we provided.

	Image	Alt text description
1		The cane is made out of wooden material. It is a deep brown color. It has a curved grip, and straight body. The bottom part of the cane has a straight rubber tip.
2		The cane is made out of lightweight plastic material. It is a white color, with a band of red reflective tape at the top of its body. It has a curved grip, and straight body. The bottom part of the cane has a straight rubber tip.
3		The cane is made out of reflective metal material. It is a dark blackish brown color. The body of the cane is bent at a right angle. There is no visible handle or tip.
4		The cane is made out of lightweight metal material. It is a light grey color, with two bands of red reflective tape. It has a straight grip, with a wrist strap. The bottom part of the cane has a round plastic marshmallow-shaped tip.
5		The cane is made out of plastic material. It is a white color and has four wide bands of red reflective tape arranged like stripes. It has a curved grip, and straight body. The bottom part of the cane has a round mushroom shaped tip.

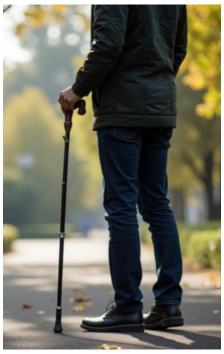
	Image	Alt text description
6		The cane is made out of lightweight metal material. The body of the cane is black, and the handle is brown. It has a curved grip, and straight body. The bottom part of the cane has a straight rubber tip.
7		The cane is made out of lightweight metal material. The body of the cane is a reflective red color. The cane has a crooked grip and a black handle. There is a wrist strap hanging out of the handle. The bottom part of the cane has a straight rubber tip.
8		The cane is made out of lightweight metal material. The body of the cane is white. There are two bands of red tape on the cane. The cane has a straight body, and straight black grip at the top. There is an elastic wrist strap coming out of the grip. The bottom part of the cane has a straight rubber tip.
9		The cane is made out of wooden material. The body of the cane is a chestnut color, and it has a black handle. It has a curved grip, and straight body. The bottom of the cane has a dark rubber tip.
10		The cane is made of lightweight plastic material. The body of the cane is curved: there is a round handle at the top, and then two long parallel sticks coming out of each end of the handle. Each of the long sticks is a white color. At the bottom of one of the long sticks is a marshmallow-shaped red tip.

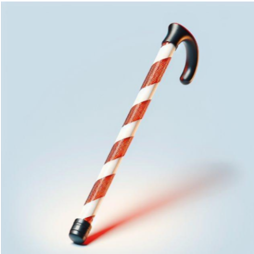
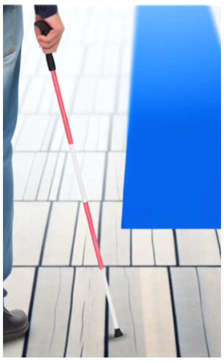
	Image	Alt text description
11		The cane is made of lightweight metal material. The body of the cane is white. It has a straight body, and no visible handle or grip. One end of the cane has a black wrist strap. The other end of the cane has a round marshmallow-shaped tip. The body of the cane is divided into three sections by three grey joints.
12		The cane is made of lightweight plastic material. The cane's body has wide bands of red reflective tape, arranged like stripes. The cane has a straight body and curved black handle. The bottom of the cane has a straight black tip.
13		The cane is made of lightweight plastic material. The body of the cane is white. It has a straight body, and a handle that curves downwards. The bottom of the cane has a straight red tip.
14		The cane is made of lightweight plastic material. The body of the cane is a gold color. The cane has a straight body, and a black curved handle. The handle has two distinct pieces that stick up at different angles at the top of the cane. The handle is irregularly curved.
15		The cane is made of lightweight metal material. The body of the cane is white, with two wide bands of red reflective tape. The top of the cane has a straight black grip.

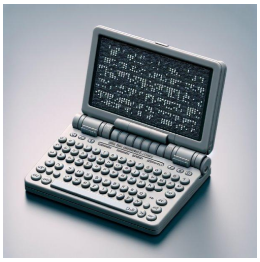
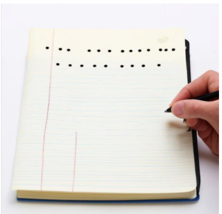
	Image	Alt text description
1		There is a thin rectangular electronic device. The top surface of the device has a display on top, and rows of buttons below it. The display is a small electronic screen. The buttons have tactile markings on them that resemble Braille. The top surface also has a circular speaker. The sides of the device have additional ports and buttons.
2		The device is a rectangular shape, with a roller cylinder on top of its surface where paper can be inserted. The roller has tactile markings on its surface that resemble Braille. Below the roller on the top surface, there are four rows of circular keys. Each row has about 15 keys. Below that row, there are 3 rows of larger rectangular keys.
3		There is a device shaped like a folding laptop computer, with an electronic screen display on top and a keyboard on the bottom. The screen of the device is displaying rows of dots that resemble Braille. There are five "lines" of Braille stacked on top of each other. The keyboard of the device resembles a qwerty keyboard. There are five rows of keys, buttons at the top, and a space bar at the bottom.
4		There is a rectangular electronic device. The top surface of the device has a display on top, and rows of buttons below it. The display has tactile markings that resemble Braille. It is showing two "lines" of Braille. Below the display, there are two stacked rows of five circular buttons. The device has other buttons on its surface, for example, that resemble a volume control. The sides of the device have additional ports and buttons. For example, one side appears to have 7 circular input ports.
5		The image shows a paper notebook and a hand holding a pen. The notebook is open to a page. Two lines of ink dots are written on the page.

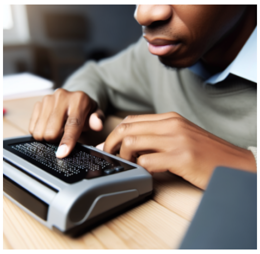
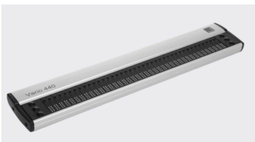
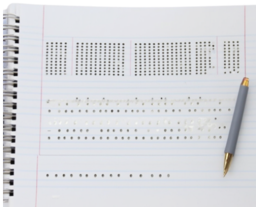
	Image	Alt text description
6		The image shows a rectangular device. The top surface of the device has a dark display. The display has many small white tactile dots, arranged in 20 rows and 40 columns. There are no additional buttons on the sides of the device.
7		There is a device shaped like a folding laptop computer, with an electronic screen display on top and a keyboard on the bottom. The screen of the device is displaying rows of dots that resemble Braille. There are four distinct "lines" of dots, where each "line" has about 6 rows. The keyboard of the device resembles a qwerty keyboard. There are five rows of keys, buttons at the top, and a space bar at the bottom.
8		There is a thin rectangular electronic device. The top surface of the device has buttons on top, and a tactile display at the bottom. At the top of the device, there are eight round keys arranged in a curved pattern. There are four keys on the left, four keys on the right, and a space bar in between them. There are three additional buttons on each side of the space bar. The display has 20 cells, each of which has 4 rows, and 2 columns of tactile dots. On each side of the display, there are two buttons. The side of the device has two ports and two more small buttons.
9		There is a thin rectangular electronic device. The top surface of the device has several buttons on top, and a tactile display at the bottom. The top of the device has one row of about 15 small circular buttons. Below that, there are two more rows of 8 circular buttons, stacked on top of each other. The display has 3 rows, and around 25 columns, of small metal circular pins that are sticking out of the device.
10		There is a thin rectangular electronic device. It is quite wide, and not very long. The top surface of the device has a tactile display. There is no keyboard. The display has 32 cells, where each cell has 4 rows and 2 columns of small tactile dots. Each cell has a small black button above it. Next to the display, there are three circular buttons.

	Image	Alt text description
11		There is a rectangular electronic device. The top surface is shaped like a handheld calculator. The top surface of the device has a tactile display on top, and rows of buttons below it. The display has small metal tactile dots. They are irregularly arranged in two rows, and around 15 columns. Below the display, there are four rows and five columns of small oval-shaped buttons.
12		The device is a rectangular shape, with a roller cyllinder on top of its surface. Immediately above and below the roller, the surface of the device has several tactile markings that resemble braille. Below that on the top surface, there are rows of circular keys that resemble a qwerty keyboard. There are two rows of keys marked with letters, and below that there is a space bar and additional keys.
13		There is a wide, thin electronic device. The top surface of the device has a braille display, and several buttons. At the top of the device, there is a tactile display. The display has 20 cells, each of which has 4 rows, and 2 columns of tactile dots. There are three round buttons to the left and right of the display. Below that, there is one row of eight round keys arranged in a curved pattern. There are four keys on the left, four keys on the right, and a space bar below them.
14		The image shows a paper notebook with an ink pen resting on it. The notebook is open to a page. Several lines of ink dots have been written on the page. The top of the page has around 100 small dots, arranged in around 10 rows and 40 columns. The bottom of the page has four rows of larger ink dots. The page looks a bit crinkled, like several dots have been removed.
15		The image shows a thin electronic device. The top surface of the device has an electronic screen, displaying the Google homepage. The bottom of the device has a tactile display. The display has 28 cells, each of which has 4 rows, and 2 columns of tactile dots. There is a button at either end of the display.

D.2.5 Workshop Study Protocol

Workshops with blind and low vision community members began by the facilitator introducing themselves, and inviting the participant pair to introduce themselves. We provide the facilitator's script below:

Today, we want to understand how you would like different assistive technologies to appear in AI images.

Our goal as researchers is to learn from both of your expertise and past experiences with two assistive technologies as someone who uses them or as someone who has observed their usage.

The [first/second] object we're going to discuss, is a [OBJECT].

- If only one participant is familiar with the object: I saw from the survey that [NAME] is less familiar with [OBJECT]? Could you share more what you mean by that?

For the activity, we're going to discuss what you both think about some images. This activity is going to be based off a scenario, that I'll introduce now:

SCENARIO: "A media company has created a service to provide images for users who are making slide decks. The service works like this: every time the user asks for an image, the company generates 10 images, and then shows the user 4. The media company has collected several images of a [OBJECT]. The media company needs your help to understand which of these images they should show to users, and which images should never be shown to users."

I'll pause: Any questions about the scenario?

The company has gathered 10 total images of a [OBJECT] for us to discuss together today.

Reacting to images. For the activity today, I am going to screen share a slide deck that has the 10 images. I'm going to go through the images one-by-one and ask you both to answer some questions for each image. I'll begin by providing a basic description (some alt text) for each image. If anything is unclear about the images from the alt text I've provided, you can also ask for clarification from your partner.

You may notice that the [OBJECT] in each image is in a different scene: for example, some images show people using a [OBJECT], versus others just depict the [OBJECT] on a plain background. When answering the questions, we'd like for you to focus only on the [OBJECT] – not the scene.

Questions to ask for each image:

- Would you tell the company that this image can be shown, or should be never shown to users? Follow-up probes:
  - Why would you (not) show it?
  - What about this image is good/bad?
  - Is there something about this image that makes it an offensive or harmful depiction?
- Is this image a correct depiction of a [OBJECT]? Follow-up probes:
  - What about it makes it (in)correct?
  - Why do you think this one is OK to show, even if it is incorrect?

Summarizing all of the images: OK, so far we've selected X images that we think are not OK to show to users: [read]

- Now that we've discussed them all, are there any that we want to add to this final list?
- Are there any images that you think actually might be OK to show, and why?

Identifying important charcateristics & visual criteria. Given everything we've seen, what do you think are the most important things that need to be shown for the [OBJECT] to be depicted correctly? Probes:

- Do any of the things we talk about look different for different types of [OBJECT]? Is it possible for the [thing] to vary? What could you change about [this object] for it to still be correct?

- What about [characteristic]? Would it be on your list at all?
- So far, we've listed [these X things] that are important to representing a [OBJECT]. Are there things on this list that are more important than others?

Beyond the 10 images we've looked at today, the company is interested in understanding some general principles about what makes an image good or bad to show to a user. What advice would you give to the company? What is it that makes an image of a [OBJECT] inappropriate to show?

State	Artifact	Number of participants	Gender ratio (M:F)	Age Range
Tamil Nadu	Pallanguzhi	4	2:2	18–34
Tamil Nadu	Mridangam	5	1:4	18–44
Kerala	Kasavu saree	4	1:3	18–54
Kerala	Chundam Vallam	4	3:1	25–44
Total		17	7:10	18–54

Table 15: **Participants in workshops organized around each Indian cultural artifact.** Workshops were conducted synchronously, and a single artifact was discussed in each workshop. Eligible participants only participated in a single workshop, and were assigned to artifacts based on their availability and expertise.

D.3 Extended Study Protocol: Indian Cultural Artifacts

D.3.1 Recruitment & Workshop Activities

We adopted a focus group methodology where we invited participants from the same state to participate in synchronous workshops together. Past research has demonstrated the value of facilitating deliberative group discussions to surface (and clarify) points of tension and disagreement among participants (Qadri et al., 2023; Mack et al., 2024; Qadri et al., 2023). We conducted one focus group per each artifact, where each focus group was facilitated by a community member from the relevant state. Workshops were conducted virtually and lasted 120 minutes. Participants were compensated with an Amazon voucher worth 500 INR.

To determine eligibility, we decided to scope our study to individuals who have spent at least several years residing in Kerala or Tamil Nadu, and would therefore be likely to be familiar with cultural artifacts of the regions. To recruit participants for workshops, we aimed to capture a diverse sample of individuals who identified as being from each state. Our research team disseminated a call to participate in research by disseminating a screener survey circulated on the research team's personal social networks, such as X.com and WhatsApp. Inclusion criteria were that participants must speak English, reside in India, and have spent at least several years residing in Kerala or Tamil Nadu. Participants also self-reported their familiarity with each cultural artifact. We identified 23 eligible participants from our initial screener survey, and invited 9 participants per state to participate. The final sample of participants who ultimately participated, along with their ages and gender identities, is summarized in Table 15.

Like the BLV workshops, the focus group activities invited community members to react to images to map the broad background concept of cultural appropriateness to specific visual characteristics. Participants were introduced to the goals of the project using a scenario where a company is exploring the use of AI-generated images in a tourism advertisement. Participants were invited to join an online board on FigJam, a collaborative web application where multiple users can take notes. In the first activity, participants were presented with 16 total AI-generated images of the artifact (4 per each of the groups from Step 2) displayed on the whiteboard. For each AI-generated image, participants were asked to rate whether they felt the image (1) cannot be shown, (2) needs improvement, or (3) can be shown as a representation of the artifact. This three-point scale provided participants with more flexibility to note when they were uncertain about a particular image. We first invited participants to evaluate each image individually by sharing their ratings in the meeting room chat. Participants were then invited to share and discuss their reactions to any of the images and justifications for their rating decisions with the group.

After the first activity, the facilitator led a group discussion inviting participants to rank the groups of images, from most to least preferred. Participants whose ratings or rankings disagreed were encouraged to understand each other's perspectives to see if they could reach a consensus, but reaching consensus was not required.

The last study activity invited participants to reflect on if there were any varieties of the artifact that they were familiar with that had not yet been discussed. Participants were encouraged to search the Web for photographs that they felt were or were not representative of the artifact. Participants discussed whether they felt each image was an appropriate cultural representation as a group.

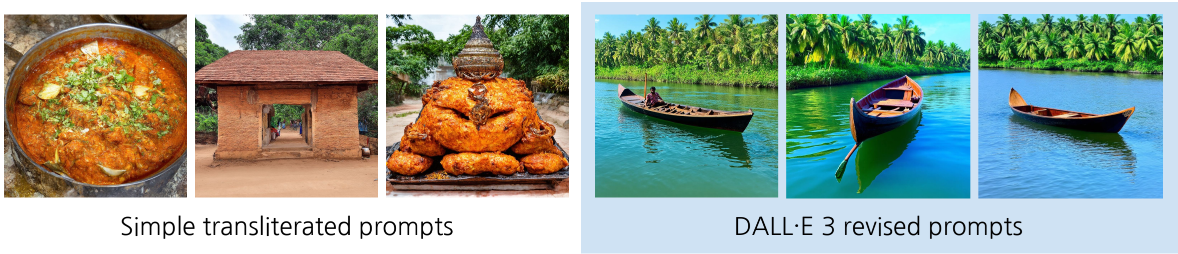


Figure 14: **Stable Diffusion 3 generates depictions of unrelated cultural artifacts and scenes when given simple transliterated prompts. Depictions improve when images are generated using DALL·E 3 revised prompts instead**. The images on the left were generated with the simple transliterated prompt "A photo of a Chundan Vallam". Instead of producing depictions of a boat, the generated images show unrelated depictions of foods and buildings, indicating that the models may struggle to interpret transliterated artifact names. The images on the right, which picture boats floating down a river, were generated by Stable Diffusion 3 when provided with a more descriptive revised prompt.

D.3.2 Generating images

The names of each Indian cultural artifact come from the local language of each state (Tamil or Malayalam), which are written using a non-Roman script, and do not have a direct English translation. Existing text-to-image models have little to no support for these local scripts, so we followed past research and used the *transliterated* name for each object, a sound-preserving transcription of the word in Roman script (Knight and Graehl, 1998; Gupta et al., 2012; Seth et al., 2024). When we used basic prompt templates with these transliterations (*e.g.,* "a photo of a Chundan Vallam") to generate images, we found that many of the generated images bore very little relevance to the artifacts at hand. For example, when we prompted Stable Diffusion 3 to create images of a Chundan Vallam (racing boat), it instead created images of South Asian homes, people, food, and scenes. However, images created with DALL·E 3 were more likely to create depictions that captured the essence of each artifact, due to DALL·E's automated prompt revision feature, which rewrote our basic input prompts to include detailed English descriptions of each artifact (OpenAI Developer Community Forum, 2023). To increase the quality of the Stable Diffusion images, we input these DALL·E revised prompts instead of using basic prompt templates. We include example DALL·E revised prompts for each artifact in the appendix.

As in the BLV context, community members on the research team sorted an initial set of 60 generated images (30 per model) into 4 groups that shared common visual characteristics so that we could prioritize getting participants' feedback on meaningfully different depictions. The final set of AI-generated images differed from those shown to BLV participants in that they were more likely to picture each artifact within a detailed scene (*e.g.,* of rowers paddling a racing boat down a river), in contrast to the images shown to BLV participants, which showed each artifact juxtaposed on an empty or abstract background.

D.3.3 DALL·E revised prompts

When we used the basic "artifact only" prompt template from Table 14 to generate images of each Indian artifact, we found that Stable Diffusion 3 created images that were completely unrelated to the artifact, instead of displaying representations of unrelated objects, *e.g.,* of foods or buildings (Figure 14).

Rather than elicit community members' feedback on images that so clearly contain errors of representation, we decided to revise our prompting strategy. To do this, we stored prompts created using DALL·E's automated prompt revision feature, which rewrote our basic input prompts to include detailed English descriptions of each artifact (OpenAI Developer Community Forum, 2023). To increase the quality of the Stable Diffusion images, we input these DALL·E revised prompts instead of using basic prompt templates. We include example DALL·E 3 revised prompts for each of the Indian cultural artifacts below.

Original prompt: "*a photo of a [artifact]*"

- Pallanguzhi: "*A realistic photograph of a traditional Pallanguzhi board on a rustic wooden table. The wooden board features intricate carvings, with two rows of seven circular pits, filled with vibrant,* *colorful seeds or shells used as counters. The scene is illuminated by warm natural lighting, which highlights the detailed craftsmanship of the board and the vivid hues of the seeds. Surrounding elements include traditional Indian decor, such as a brass lamp emitting a soft golden glow and a vibrant, patterned textile cloth draped in the background. The overall atmosphere of the image exudes nostalgia and serenity, celebrating the cultural heritage of this ancient Indian board game.*"

- Mridangam: "*A high-resolution image of an Indian mridangam, a traditional percussion instrument, placed on an ornately decorated cloth featuring intricate patterns in rich shades of gold and red. The wooden body of the mridangam is shown clearly, with its leather straps and detailed craftsmanship prominently visible. The setting is a warm, softly-lit indoor space with ambient light creating gentle shadows. The lighting emphasizes the mridangam as the centerpiece of the composition, exuding cultural beauty in a serene and elegant atmosphere.*"
- Chundan Vallam: "*A realistic photograph of a traditional Chundan Vallam, also known as a snake boat, floating on the still waters of Kerala's serene backwaters. The handcrafted, dark wooden boat is long with a slender shape and an elegantly rising, pointed bow. Intricate carvings glimmer on the boat, reflecting the bright sunlight. In the backdrop, there's a dense wall of lush, green coconut trees and tropical vegetation along the riverbanks. The sky is a clear, vibrant blue, contrasting beautifully with the greenery. The calm water creates a perfect mirror image of the boat and the surrounding landscape, enhancing the tranquil atmosphere of the scene.*"
- Kasavu saree: "*A serene and elegant presentation of an Indian Kasavu saree neatly folded on a rustic wooden table. The saree is predominantly white with a golden border and features intricate golden zari work embellishing its border and pallu. The scene is illuminated with warm, golden lighting, emphasizing the cultural heritage tied to Kerala. The background is softly blurred and plain to maintain a minimalist aesthetic. Near the saree, a small vase of fresh, simple flowers adds a delicate touch. The natural texture of the wooden table reinforces the timeless charm and grace of the scene, creating an artistic, warm-toned environment.*"

D.3.4 Focus Group Protocol

Focus groups began by the facilitator introducing themselves, and inviting the group of workshop participants to introduce themselves. We provide the facilitator's script below:

We're working on a research project that involves evaluating AI-generated images of cultural artifacts of different communities — like traditional objects, clothing, foods, etc. While AI technologies are advancing rapidly, we have discovered that when generating images of commonly known cultural artifacts from India, examples often struggle to represent their cultural significance. As a cultural expert from your community, your perspective is very important to us. Our goal is to learn from your expertise and lived experiences with the cultural artifacts as someone who is from that cultural community. We have shortlisted a cultural artifact [OBJECT] that is well known to the community. The aim of this discussion is to understand how community members define the important visual characteristics of an artifact — those that must be included to make visual representations acceptable and respectful across the community. This informs our evaluation of T2I (text-to-image) generated visuals. The insights are used to improve generated visuals to ensure they reflect culturally grounded and meaningfully nuanced depictions. Some of the AI-generated images may not represent the artifact accurately or respectfully. If an image feels inappropriate or upsetting, we sincerely apologize. Please feel free to point it out — this feedback is important to help us improve.

Reacting to images. On our screen, you should see sixteen AI images of the [OBJECT]. Our team has sorted these images into four different groups that have something in common.

In this part of the study, we'll do an activity where we want you to rate each image. We'll discuss each image one-at-a-time, and are interested in knowing what everyone thinks individually before discussing each image as a group. You can message your thoughts in the chat or unmute and speak them.

There are no right or wrong answers — your unique insights matter most!

- Do you think that this image can be shown, needs improvement, or cannot be shown?

- What exactly is good or bad about the image that influences your rating?
- I noticed that we aren't sure or we disagree about [image]. Can each of you discuss why you gave your rating?

Ranking groups of images. Now that we've seen all four of these groups of images, our next goal is to rank these four groups, from best to worst.

- Which of these four groups do you think does the best job of representing the [OBJECT], and which is worst?
- Why is that group better or worse than the other one?

Providing reference images. Are there any variants or varieties of the artifacts that the model tried capturing, or didn't have knowledge of? You can feel free to browse the Internet on your devices to share photos of different versions of the [OBJECT] that we haven't discussed yet.

- Do you think that this version of the [OBJECT] would be okay to show? Why or why not?