

# Community-Centered Measurement of Cultural Content in AI Images

Anonymous ACL submission

## 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.

## 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; Exner et al., 2025). 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; Kunchukuttan et al., 2020; Seth et al., 2024; Magomere et al., 2025; Romero et al., 2024) 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 human 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

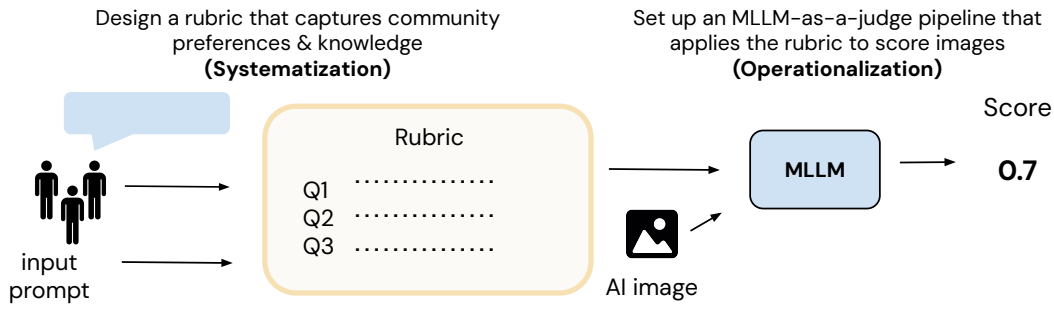


Figure 1: **Scaffolding community engagement to develop community-centered measures of cultural representation.** Given an input prompt (e.g., “a photo of a guide cane”), we invited community members to participate in designing a rubric that captures their expertise and preferences for each cultural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

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 on data 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

155 sharing learnings and opportunities for future re-  
156 search to bring community expertise into the design  
157 of AI evaluation metrics.

## 158 2 Related Work

159 **Cultural Representation and AI-Generated Me-**  
160 **dia.** AI-generated images have the power to shape  
161 how communities are understood and perceived  
162 by others, as well as how they perceive them-  
163 selves (Qadri et al., 2023; Hall, 1997). Marginal-  
164 ized communities have used AI tools to create vi-  
165 sual media that aligns with their goals for represen-  
166 tation (Das et al., 2024; Huh et al., 2023; Adnin and  
167 Das, 2024; He et al., 2025; Hamna et al., 2025b).  
168 However, these same AI tools have been shown  
169 to reproduce normative identities and narratives  
170 that contribute to the erasure of already marginal-  
171 ized communities (Bianchi et al., 2023; Gillespie,  
172 2024; Mack et al., 2024), resulting in representa-  
173 tional harm. In this paper, we focus on the issue  
174 of appropriate and desired cultural representation—  
175 an often contested concept (Chasalow and Levy,  
176 2021). We adopt a broad conceptualization of cul-  
177 ture as based on shared identity, recognizing that  
178 individuals can inhabit multiple cultural identities  
179 simultaneously (Zhou et al., 2025). Culture may  
180 be grounded in place (*e.g.*, shared nationality (Adi-  
181 lazuarda et al., 2024)), or in other identity aspects  
182 including race or ethnicity (Egede, 2025), disabili-  
183 ty (Mack et al., 2024), sexuality (Taylor et al.,  
184 2024), profession (Tseng et al., 2025), or relational  
185 roles (Viswanathan et al., 2025).

186 In this paper, we focus on designing quantitative  
187 measures that reflect participants’ lived understand-  
188 ings of their culture. Prior work in AI and culture  
189 has emphasized the need for methodologies that  
190 “translate qualitative insights into algorithmic inter-  
191 ventions” (Biega et al., 2025). Our work responds  
192 to this call by bringing qualitative methods into the  
193 design of quantitative metrics by engaging commu-  
194 nities in the design of rubrics that can be applied  
195 with MLLM-as-a-judge systems.

196 **Measuring the Performance of Generative Text-**  
197 **to-Image Models.** Our work contributes to a grow-  
198 ing body of interdisciplinary scholarship focused  
199 on developing and critiquing performance mea-  
200 sures used to evaluate generative AI systems. De-  
201 spite the recognized importance of evaluation in  
202 generative AI development, significant challenges  
203 remain. Given the speed at which the AI indus-  
204 try is moving, there is a need for evaluation prac-

205 tices to be repeatable, automatable, and applica-  
206 ble across different models, datasets, and settings  
207 (Shankar et al., 2024; Weidinger et al., 2023; Har-  
208 vey et al., 2025). It is increasingly common to auto-  
209 mate evaluations using model-mediated evaluation  
210 approaches, where an auxiliary, pretrained model’s  
211 internal representations or judgments are used to  
212 stand in for human evaluative judgments as a form  
213 of “ground truth” (Kapania et al., 2025). Drawing  
214 on the affordances for scale of LLM-as-a-judge  
215 pipelines, practitioners have increasingly turned to  
216 multimodal language models (MLLMs) as judges  
217 for evaluating generated images (Hu et al., 2023;  
218 Zhang et al., 2023; Lin et al., 2024; Chen et al.,  
219 2025). In these pipelines, a multimodal model  
220 scores each image using a rubric, which in some  
221 cases is generated by another model. Such ap-  
222 proaches, however, introduce a fundamental depen-  
223 dency: the quality of the measurements becomes  
224 limited by the capabilities and biases of the auxil-  
225 iary models (Magomere et al., 2025; Kreiss et al.,  
226 2022; Seth et al., 2024; Kapania et al., 2025; Mas-  
227 siceti et al., 2024). Since these models are trained  
228 on historical datasets that may lack diversity or  
229 contain societal biases, they can systematically un-  
230 dervalue or misrepresent content from marginal-  
231 ized communities. For example, if an auxiliary  
232 model was trained primarily on Western imagery,  
233 it may poorly evaluate AI-generated images of non-  
234 Western cultural practices, foods, or aesthetics. We  
235 ask whether this limitation can be overcome by the  
236 design of more thoughtful rubrics.

## 237 3 Methods

238 **Selecting communities and cultural artifacts.**  
239 We collaborated with members of three distinct  
240 communities: blind and low-vision (BLV) people  
241 located in the United Kingdom, and current and for-  
242 mer residents of two states in South India: Tamil  
243 Nadu and Kerala. We chose these specific com-  
244 munities because members of the research team  
245 belong to them. This close researcher–community  
246 collaboration supported activities such as artifact  
247 selection, participant recruitment, and interpreta-  
248 tion of our findings.

249 For the BLV community, we focused our study  
250 on a single region (the UK) where our authorship  
251 team was based. We chose to geographically situ-  
252 ate our study in the UK because the material cul-  
253 ture of blindness is shaped by local factors, such  
254 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 C.2.1.

For each artifact, we curated a small set of AI-generated images and real photographs to

321 facilitate discussions of cultural representation.  
322 We generated images using two state-of-the-art  
323 text-to-image models at the time of the study,  
324 Stable Diffusion 3 (Stability AI, 2024a) and  
325 DALL·E 3 (OpenAI, 2024), prompting each model  
326 with a simple template (e.g., “A photo of a {artifact  
327 name}”). For each artifact, we generated an initial  
328 pool of images and then grouped images by shared  
329 visual characteristics, sampling across groups to  
330 create a manageable but meaningfully diverse  
331 set for participants to review (Mack et al., 2023).  
332 The final set of images used in the workshops is  
333 included in Appendix C.

334 For each artifact, we conducted two study activi-  
335 ties. First, after presenting each image (by screen-  
336 sharing and/or presenting an alt-text description for  
337 BLV participants), the facilitator asked participants  
338 to share if they felt that the image could or could  
339 not be shown to the general public to represent the  
340 artifact and why. The facilitator then prompted  
341 each participant to elaborate on what made a par-  
342 ticular image a good, bad, offensive, or incorrect  
343 depiction of the artifact.

344 Second, we invited participants to reflect on what  
345 they had seen so far to create a list of the most  
346 important criteria for a culturally appropriate por-  
347 trayal of the object and provide reasons for each  
348 of their responses. This activity encouraged partic-  
349 ipants to articulate concrete visual criteria that  
350 shaped their decisions. The study facilitator en-  
351 couraged participants to reflect on the degree of  
352 acceptable variability for objects and how visual  
353 features should be prioritized.

354 **Systematizing cultural representation into**  
355 **rubrics.** We used the data collected from work-  
356 shops to develop systematized concepts of “cultural  
357 appropriateness” for each artifact. These took the  
358 form of simple rubrics made up of binary yes/no  
359 criteria that community members identified as inte-  
360 gral to an accurate and appropriate representation  
361 of the artifact; see Figure 3 for an example of the  
362 rubric produced for a guide cane.

363 To distill criteria for the rubrics, we first orga-  
364 nized participants’ statements from both activities  
365 into two categories: (1) *criteria*, or the concrete vi-  
366 sual features or physical characteristics that shaped  
367 participants’ decision-making, and (2) *justifica-*  
368 *tions* that participants provided for their decisions.  
369 In deciding which criteria to include in the rubrics,  
370 we made two key decisions. First, we excluded cri-  
371 teria that were highly contested across participants.

372 We also excluded criteria that participants consis-  
373 tently described as less essential to depicting the  
374 essence of each artifact. Consequently, rather than  
375 offering a complete description of each artifact, our  
376 rubrics focus on the core features that participants  
377 agreed were *most essential* to a culturally appropri-  
378 ate depiction.

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

389 Using the rubric, we say that an image is “cultur-  
390 ally appropriate” only if it satisfies all of the rubric  
391 criteria, receiving a positive label of 1. Otherwise,  
392 we say that an image is culturally inappropriate  
393 (label of 0). While these final binary cultural ap-  
394 propriateness labels do not capture the subjective  
395 and contested nature of cultural representation, we  
396 adopt this aggregation function for its simplicity to  
397 calculate, and its direct comparability with partici-  
398 pants’ judgments.

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

414 **Operationalizing the rubrics through MLLM-**  
415 **as-a-judge.** To operationalize our systematized  
416 concepts, we *explore* whether our community-  
417 elicited rubrics can be automatically evaluated us-  
418 ing an MLLM-as-a-judge pipeline. We adopt a  
419 simple system prompt, following past work (Zheng  
420 et al., 2023; Chen et al., 2025), that instructs the  
421 MLLM to “*determine whether the provided image*  
422 *meets each criterion*”. We adopt GPT 4-o (OpenAI

Themes	Criteria
THEME 1. <i>"The object needs to be functional as an assistive technology, and usable by someone who is blind."</i>	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. <i>"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."</i>	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 B.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 B.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	MLLM	Agreement	Agreement	Agreement
	(% Appropriate)	(% Appropriate)	Overall	Appropriate images	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 B.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 B.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

561 errors: low agreement on inappropriate images  
562 indicates the MLLM’s failure to recognize violated  
563 criteria, while low agreement on appropriate  
564 images (which we see for Pallanguzhi and braille  
565 notetakers) may suggest that the MLLM lacks the  
566 cultural understanding needed to recognize valid  
567 depictions. Each has different implications for  
568 how practitioners might revise their rubric criteria  
569 to improve judge performance, as the best steps  
570 forward (*e.g.*, relaxing or clarifying criteria in cases  
571 of false negatives, or strengthening or expanding  
572 criteria in cases of false positives) will depend  
573 on the type of error being made. We include  
574 additional analyses that break down agreement for  
575 each individual rubric criterion in Appendix B.3.4.

## 576 5 Discussion

577 This paper centers the perspectives of impacted  
578 communities in designing quantitative evaluations  
579 of cultural representation in AI-generated images.  
580 We provide an initial set of results to show how  
581 community expertise and perspectives can be  
582 elicited and encoded by *systematizing* criteria for  
583 cultural representation in the form of rubrics, and  
584 *operationalized* by applying those rubrics with an  
585 MLLM-as-a-judge. We hope that researchers and  
586 practitioners can extend on the methods presented  
587 to invite community members to participate in the  
588 design of evaluation metrics. In what follows, we  
589 share several lessons learned and open research  
590 directions surfaced by this exploratory work.

591 Examining the differences between community-  
592 elicited versus LLM-generated rubrics reveals  
593 the value of inviting community participation in  
594 measurement design. Rather than providing a  
595 surface-level description of the visual attributes of  
596 each artifact, participants’ descriptions spoke to  
597 how artifacts are used and experienced in practice.  
598 In contrast, many LLM-generated rubrics failed  
599 to capture these salient features and in some cases  
600 reflected confused or inaccurate understandings.  
601 This finding extends those from past work, which  
602 describe how LLM-generated rubrics are often  
603 overly vague or underspecified (Szymanski et al.,  
604 2024a), to reveal how these rubrics may reflect  
605 deeper cultural misalignment. Taken together,  
606 these results suggest that community participation  
607 in defining evaluation criteria can help bridge epis-  
608 temic gaps in large pretrained models, particularly  
609 for low-data cultural artifacts (Massiceti et al.,  
610 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 sys-  
tematization. Addressing these questions requires  
grappling with both technical challenges in the de-  
sign 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 mea-  
surement 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 bene-  
fits of human-centered approaches to measurement  
through close collaborations with three communi-  
ties and just a small set of artifacts, many open  
questions remain. In real-world settings, founda-  
tion 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 iden-  
tifying representational harms (*e.g.*, (Mack et al.,  
2023; Qadri et al., 2025)) are not feasible to ap-  
ply. 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 commu-  
nity members to assess the face validity of LLM-  
generated rubrics as a lightweight check. If the  
rubrics diverge significantly from community mem-  
bers’ 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 engage-  
ments around content with strong normative stakes  
and contested meanings, where it is important for  
community members to exercise authority in shap-  
ing how they are represented (Qadri et al., 2025).

660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716

## 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](#). *Preprint*, arXiv:2403.15412.

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.

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.

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. 717  
718  
719

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. 720  
721  
722  
723

Jiahui Chen, Candace Ross, Reyhane Askari-Hemmat, Koustuv Sinha, Melissa Hall, Michal Drozdal, and Adriana Romero-Soriano. 2025. [Multi-modal language models as text-to-image model evaluators](#). *Preprint*, arXiv:2505.00759. 724  
725  
726  
727  
728

Tim Connell. 2008. [The challenge of assistive technology and braille literacy](#). [Online; accessed 6-September-2025]. 729  
730  
731

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. 732  
733  
734  
735  
736  
737

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. 738  
739  
740  
741  
742  
743  
744

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). 745  
746  
747  
748  
749

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. 750  
751  
752  
753

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. 754  
755  
756  
757

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](#). *Preprint*, arXiv:2502.06559. 758  
759  
760  
761  
762

Yannick Exner, Jochen Hartmann, Oded Netzer, and Shunyuang Zhang. 2025. [Ai in disguise - how ai-generated ads' visual cues shape consumer perception and performance](#). 763  
764  
765  
766

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. 767  
768  
769  
770

771	Sanjana Gautam, Pranav Narayanan Venkit, and Sourojit Ghosh. 2024. From melting pots to misrepresentations: Exploring harms in Generative AI. <i>arXiv preprint arXiv:2403.10776</i> .	829
772		830
773		831
774		832
775	Simret Araya Gebreegzabher, Charles Chiang, Zichu Wang, Zahra Ashktorab, Michelle Brachman, Werner Geyer, Toby Jia-Jun Li, and Diego Gómez-Zar4. 2025. <a href="#">Metricmate: An interactive tool for generating evaluation criteria for llm-as-a-judge workflow</a> . In <i>Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work</i> , CHIWORK '25, New York, NY, USA. Association for Computing Machinery.	833
776		834
777		835
778		836
779		837
780		838
781		839
782		840
783		841
784	Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson, and Aylin Caliskan. 2024. <a href="#">Do generative ai models output harm while representing non-western cultures: Evidence from a community-centered approach</a> . <i>Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society</i> , 7(1):476–489.	842
785		843
786		844
787		845
788		846
789		847
790	Tarleton Gillespie. 2024. <a href="#">Generative ai and the politics of visibility</a> . <i>Big Data &amp; Society</i> , 11(2):20539517241252131.	848
791		849
792		850
793		851
794	Kanika Gupta, Monojit Choudhury, and Kalika Bali. 2012. <a href="#">Mining Hindi-English transliteration pairs from online Hindi lyrics</a> . In <i>Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)</i> , pages 2459–2465, Istanbul, Turkey. European Language Resources Association (ELRA).	852
795		853
796		854
797		855
798		856
799		857
800		858
801	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? <i>arXiv preprint arXiv:2309.07462</i> .	859
802		860
803		861
804		862
805		863
806	Melissa Hall, Samuel J. Bell, Candace Ross, Adina Williams, Michal Drozdal, and Adriana Romero Soriano. 2024. <a href="#">Towards geographic inclusion in the evaluation of text-to-image models</a> . In <i>Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency</i> , FAccT '24, page 585–601, New York, NY, USA. Association for Computing Machinery.	864
807		865
808		866
809		867
810		868
811		869
812		870
813		871
814	Siobhan Mackenzie Hall, Samantha Dalal, Raesetje Se-fala, Foutse Yuehgo, 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. <i>arXiv preprint arXiv:2502.05961</i> .	872
815		873
816		874
817		875
818		876
819		877
820	Stuart Hall, editor. 1997. <i>Representation: Cultural Representations and Signifying Practices</i> . Sage Publications, London.	878
821		879
822		880
823	Hamna, Gayatri Bhat, Sourabrata Mukherjee, Faisal Lalani, Evan Hadfield, Divya Siddarth, Kalika Bali, and Sunayana Sitaram. 2025a. <a href="#">Building benchmarks from the ground up: Community-centered evaluation of llms in healthcare chatbot settings</a> . <i>Preprint</i> , arXiv:2509.24506.	881
824		882
825		883
826		884
827		885
828		886
	Hamna, Deepthi Sudharsan, Agrima Seth, Ritvik Budhiraja, Deepika Khullar, Vyshak Jain, Kalika Bali, Aditya Vashistha, and Sameer Segal. 2025b. <a href="#">Kahani: Culturally-nuanced visual storytelling tool for non-western cultures</a> . In <i>Proceedings of the 2025 ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies</i> , COMPASS '25, page 379–400, New York, NY, USA. Association for Computing Machinery.	829
		830
		831
		832
		833
		834
		835
		836
		837
	Emma Harvey, Emily Sheng, Su Lin Blodgett, Alexandra Chouldechova, Jean Garcia-Gathright, Alexandra Olteanu, and Hanna Wallach. 2025. <a href="#">Understanding and meeting practitioner needs when measuring representational harms caused by llm-based systems</a> . <i>Preprint</i> , arXiv:2506.04482.	838
		839
		840
		841
		842
		843
	Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. <a href="#">LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts</a> . In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.	844
		845
		846
		847
		848
		849
		850
		851
	Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. 2025. <a href="#">Dream-story: Open-domain story visualization by llm-guided multi-subject consistent diffusion</a> . <i>IEEE Transactions on Pattern Analysis and Machine Intelligence</i> .	852
		853
		854
		855
		856
		857
		858
	Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. <a href="#">Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering</a> . <i>Preprint</i> , arXiv:2303.11897.	859
		860
		861
		862
		863
	Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. <a href="#">Genassist: Making image generation accessible</a> . In <i>Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology</i> , pages 1–17.	864
		865
		866
		867
	Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K Reddy, and Sunipa Dev. 2024. <a href="#">Visage: A global-scale analysis of visual stereotypes in text-to-image generation</a> . <i>arXiv preprint arXiv:2401.06310</i> .	868
		869
		870
		871
		872
	Harry H. Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. 2023. <a href="#">Ai art and its impact on artists</a> . In <i>Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society</i> , AIES '23, page 363–374, New York, NY, USA. Association for Computing Machinery.	873
		874
		875
		876
		877
		878
		879
	Nari Johnson, Hamna Abid, Deepthi Sudharsan, Theo Holroyd, Samantha Dalal, Siobhan Mackenzie Hall, Jennifer Wortman Vaughan, Daniela Massiceti, and Cecily Morrison. 2025. <a href="#">Position: To make text-to-image models that work for marginalized communities, we need new measurement practices for the long tail</a> .	880
		881
		882
		883
		884
		885
		886

887	Shivani Kapania, Stephanie Ballard, Alex Kessler, and Jennifer Wortman Vaughan. 2025. Examining the expanding role of synthetic data throughout the ai development pipeline. In <i>Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency</i> .	<i>the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 12172–12182.	944
888			945
889		J. Nathan Matias and Megan Price. 2025. <a href="#">How public involvement can improve the science of ai</a> .	946
890		<i>Proceedings of the National Academy of Sciences</i> , 122(48):e2421111122.	947
891			948
892			949
893	Kevin Knight and Jonathan Graehl. 1998. <a href="#">Machine transliteration</a> . <i>Computational Linguistics</i> , 24(4):599–612.	Annika Muehlbradt and Shaun K. Kane. 2022. <a href="#">What’s in an alt tag? exploring caption content priorities through collaborative captioning</a> . <i>ACM Trans. Access. Comput.</i> , 15(1).	950
894			951
895			952
896	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. <i>arXiv preprint arXiv:2205.10646</i> .	Peter Newmark. 1988. <i>A Textbook of Translation</i> , volume 66. Prentice Hall, New York.	953
897			954
898			955
899			956
900			957
901	Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N. C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. <a href="#">Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages</a> . <i>Preprint</i> , arXiv:2005.00085.	OpenAI. 2022. <a href="#">DALL-E now available without waitlist</a> . Accessed: 2024-05-28.	958
902			959
903		OpenAI. 2024. <a href="#">DALL-E 3</a> . <a href="https://openai.com/index/dall-e-3/">https://openai.com/index/dall-e-3/</a> . Accessed: 2025-09-11.	960
904			961
905		OpenAI. 2025. <a href="#">Gpt-image-1</a> . <a href="https://platform.openai.com/docs/guides/image-generation?image-generation-model=gpt-image-1">https://platform.openai.com/docs/guides/image-generation?image-generation-model=gpt-image-1</a> . Accessed: 2025-09-11.	962
906			963
907	Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. <a href="#">From generation to judgment: Opportunities and challenges of llm-as-a-judge</a> . <i>Preprint</i> , arXiv:2411.16594.	OpenAI and others. 2024. <a href="#">Gpt-4o system card</a> . <i>Preprint</i> , arXiv:2410.21276.	964
908			965
909		OpenAI Developer Community Forum. 2023. <a href="#">Api image generation in dall-e-3 changes my original prompt without my permission</a> . [Online; accessed 6-September-2025].	966
910			967
911			968
912			969
913	Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. <a href="#">Evaluating text-to-visual generation with image-to-text generation</a> . <i>Preprint</i> , arXiv:2404.01291.	Iro-ro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Mari-vate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Ar-shath Ramkilowan, Adewale Akinfaderin, Alp Ök-tem, Wole Akin, Ghollah Kioko, and 6 others. 2020. <a href="#">Masakhane – machine translation for africa</a> . <i>Preprint</i> , arXiv:2003.11529.	970
914			971
915			972
916			973
917			974
918	Kelly Mack, Rai Ching Ling Hsu, Andrés Monroy-Hernández, Brian A. Smith, and Fannie Liu. 2023. <a href="#">Towards inclusive avatars: Disability representation in avatar platforms</a> . In <i>Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems</i> , CHI ’23, New York, NY, USA. Association for Computing Machinery.		975
919			976
920			977
921			978
922			979
923			980
924			981
925	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 <i>Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems</i> , pages 1–23.	Lawrence Palinkas, Sarah Horwitz, Carla Green, Jennifer Wisdom, Naihua Duan, and Kimberly Hoag-wood. 2013. <a href="#">Purposeful sampling for qualitative data collection and analysis in mixed method implementation research</a> . <i>Administration and policy in mental health</i> , 42.	982
926			983
927			984
928			985
929			986
930			987
931	Jabez Magomere, Shu Ishida, Tejumade Afonja, Aya Salama, Daniel Kochin, Yuehgo 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 <i>Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency</i> , pages 246–282.	Rida Qadri, Mark Diaz, Ding Wang, and Michael Madaio. 2025. The Case for "Thick Evaluations" of Cultural Representation in AI. <i>arXiv preprint arXiv:2503.19075</i> .	988
932			989
933			990
934			991
935			992
936			993
937			994
938			995
939			996
940	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 <i>Proceedings of</i>	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 <i>Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency</i> , pages 506–517.	997
941			998
942			999
943			1000

995	David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. <a href="#">Cvqa: Culturally-diverse multilingual visual question answering benchmark</a> . <i>Preprint</i> , arXiv:2406.05967.	1052
996		1053
997		1054
998		1055
999		1056
1000		1057
1001		1058
1002		
1003	Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. Dosa: A dataset of social artifacts from different indian geographical subcultures. <i>arXiv preprint arXiv:2403.14651</i> .	
1004		
1005		1059
1006		1060
1007		1061
1008		1062
1009	Shreya Shankar, J.D. Zamfirescu-Pereira, Bjoern Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. <a href="#">Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences</a> . In <i>Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology</i> , UIST '24, New York, NY, USA. Association for Computing Machinery.	1063
1010		1064
1011		1065
1012		1066
1013		
1014		1067
1015		1068
1016		1069
1017	Devinder Singh and Manoj Sharma. 2009. <a href="#">Unfolding the indian cultural mosaic: a cross-cultural study of four regional cultures</a> . <i>International Journal of Indian Culture and Business Management - Int J Indian Cult Bus Manag</i> , 2.	1070
1018		1071
1019		1072
1020		1073
1021		1074
1022	Stability AI. 2024a. <a href="#">Announcing the open release of stable diffusion 3 medium, our most sophisticated image generation model to date</a> . Accessed: 2025-09-11.	
1023		
1024		
1025		
1026	Stability AI. 2024b. <a href="#">Introducing stable diffusion 3.5</a> . Accessed: 2025-09-11.	
1027		
1028		1075
1029	Harini Suresh, Divya Shanmugam, Tiffany Chen, Annie G Bryan, Alexander D'Amour, John Guttag, and Arvind Satyanarayan. 2023. <a href="#">Kaleidoscope: Semantically-grounded, context-specific ml model evaluation</a> . In <i>Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems</i> , CHI '23, New York, NY, USA. Association for Computing Machinery.	1076
1030		1077
1031		1078
1032		1079
1033		1080
1034		1081
1035		1082
1036	Annalisa Szymanski, Simret Araya Gebreegziabher, Oghenemaro Anuyah, Ronald A. Metoyer, and Toby Jia-Jun Li. 2024a. <a href="#">Comparing criteria development across domain experts, lay users, and models in large language model evaluation</a> . <i>Preprint</i> , arXiv:2410.02054.	1083
1037		1084
1038		1085
1039		
1040		1086
1041		1087
1042	Annalisa Szymanski, Noah Ziemis, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. 2024b. <a href="#">Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks</a> . <i>Preprint</i> , arXiv:2410.20266.	1088
1043		1089
1044		1090
1045		1091
1046		
1047	S. J. Tambiah. 1967. <a href="#">Social change in modern india</a> . by m. n. srinivas. university of california press: Berkeley and los angeles, and cambridge university press: London, 1966. pp. xv + 194, 40s. <i>Modern Asian Studies</i> , 1(4):404–405.	1092
1048		1093
1049		1094
1050		1095
1051		1096
	Jordan Taylor, Ellen Simpson, Anh-Ton Tran, Jed R. Brubaker, Sarah E Fox, and Haiyi Zhu. 2024. <a href="#">Cruising queer hci on the dl: A literature review of lgbtq+ people in hci</a> . In <i>Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems</i> , CHI '24, New York, NY, USA. Association for Computing Machinery.	1097
		1098
		1099
		1100
		1101
	Emily Tseng, Meg Young, Marianne Aubin Le Quéré, Aimee Rinehart, and Harini Suresh. 2025. <a href="#">"ownership, not just happy talk": Co-designing a participatory large language model for journalism</a> . In <i>Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency</i> , FAccT '25, page 3119–3130, New York, NY, USA. Association for Computing Machinery.	1102
		1103
		1104
		1105
		1106
	Sruthi Viswanathan, Seray Ibrahim, Ravi Shankar, Reuben Binns, Max Van Kleek, and Petr Slovak. 2025. <a href="#">The interaction layer: An exploration for co-designing user-llm interactions in parental wellbeing support systems</a> . In <i>Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems</i> , CHI '25, New York, NY, USA. Association for Computing Machinery.	1107
		1108
	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 <i>Proceedings of the 42nd International Conference on Machine Learning (ICML)</i> .	
	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. <i>arXiv preprint arXiv:2310.11986</i> .	
	Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. <a href="#">Benchmarking machine translation with cultural awareness</a> . In <i>Findings of the Association for Computational Linguistics: EMNLP 2024</i> , pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.	
	Meg Young. 2025. <a href="#">Participatory ai? begin with the most affected people</a> . <i>TechPolicy.Press</i> . 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. <i>First Monday</i> .	
	Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang	

- 1109 Wang, and Linda Ruth Petzold. 2023. [Gpt-4v\(ision\)](#)  
1110 [as a generalist evaluator for vision-language tasks.](#)  
1111 *Preprint*, arXiv:2311.01361.
- 1112 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan  
1113 Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,  
1114 Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang,  
1115 Joseph E. Gonzalez, and Ion Stoica. 2023. [Judg-](#)  
1116 [ing llm-as-a-judge with mt-bench and chatbot arena.](#)  
1117 *Preprint*, arXiv:2306.05685.
- 1118 Naitian Zhou, David Bamman, and Isaac L Bleaman.  
1119 2025. Culture is not trivia: Sociocultural theory for  
1120 cultural nlp. *arXiv preprint arXiv:2502.12057*.

## A Systematized Concepts

1122

1123

1124

1125

1126

1127

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.

<i>Theme</i>	<i>Filter-out criteria for guide cane</i>	<i>Filter-out criteria for braille notetaker</i>
<p><b>Theme 1.</b> The object needs to be functional as an assistive technology, and usable by someone who is blind.</p>	<p><b>C1.</b> No deformed canes.  <b>C2.</b> No curved (crooked) handles.  <b>C3.</b> The cane must be shaped like a long (5 foot) stick.  <b>C4.</b> The body must have sections that are a white color.  <b>C5.</b> There must be a tip at the bottom of the cane.</p>	<p><b>C1.</b> The device must be shaped like a thin rectangular box.</p> <p>The device must have a valid braille output, to read:  <b>C2.</b> The device must show braille.  <b>C3.</b> All depictions of braille must be tactile (embossed). No depictions of braille on electronic screens.  <b>C4.</b> Depictions of braille must be valid: arranged in cells with 3 or 4 rows, and 2 columns.</p> <p>The device must have a valid braille input, to write:  <b>C5.</b> 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.</p>
<p><b>Theme 2.</b> 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.</p>	<p><b>C1.</b> No wooden walking sticks.  <b>C2.</b> No decorative striped patterns (candy canes).</p>	<p><b>C1.</b> No depictions of notetaking as writing using a pen on paper.  <b>C2.</b> No devices that are shaped like laptops with an electronic screen output.  <b>C3.</b> No devices that are shaped like handheld calculators with an electronic screen output.  <b>C4.</b> No devices that are shaped like manual typewriters.</p>

Table 2: **Filter-out criteria for depictions of a guide cane and braille notetaker.** The criteria are organized under two themes. An image is defined to be culturally inappropriate if at least one criterion is not met, and culturally appropriate otherwise.

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

**Table 3: Filter-out criteria for depictions of a Chundan Vallam, Mridangam, Pallanguzhi, and Kasavu saree.** The criteria for each artifact are organized under three themes. An image is defined to be culturally inappropriate if at least one criterion is not met, and culturally appropriate otherwise.

## B Experiments & Results: Extended

### B.1 RQ1: Systematizing community expertise into evaluation rubrics

#### B.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.

Correcting errors Braille notetaker		Clarifying criteria Braille notetaker		Adding criteria Mridangam	
LLM	Includes a screen or output display for visual feedback.	LLM	Displays rows of raised dots for braille reading.	LLM	...
Ours	No devices with an electronic screen output.	Ours	The braille must be arranged in valid cells: dots in 3 or 4 rows, 2 columns.	Ours	A black circular membrane must be present on both drumheads.

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 B.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.

### B.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.

## B.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 notetaking 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 boards 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).

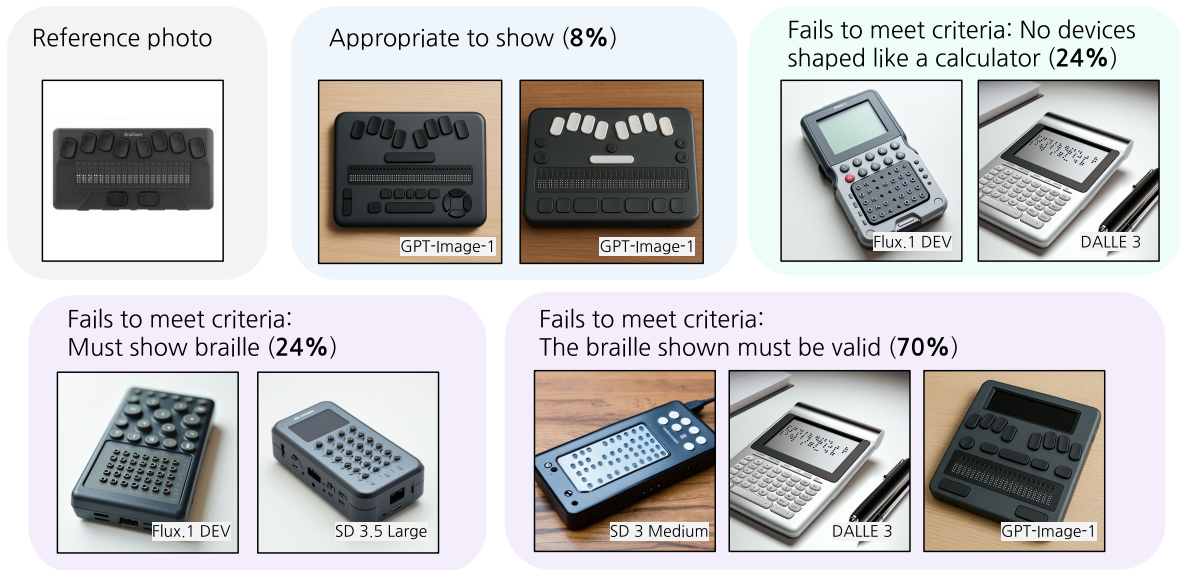


Figure 11: **Criterion-level annotations provided by humans reveal the specific representational errors that make depictions of a braille notetaker inappropriate.** The figure displays a reference photo of a braille notetaker, and example AI-generated images that fall into one of four groups (as annotated 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.

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

#### B.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}”<sup>1</sup>. 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 13). For the Indian cultural artifacts, we used revised DALL·E 3 prompts as detailed in Appendix C.3.3.

#### B.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
Braille 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).

<sup>1</sup>DALL·E 3 automatically enriches this prompt with more detail.

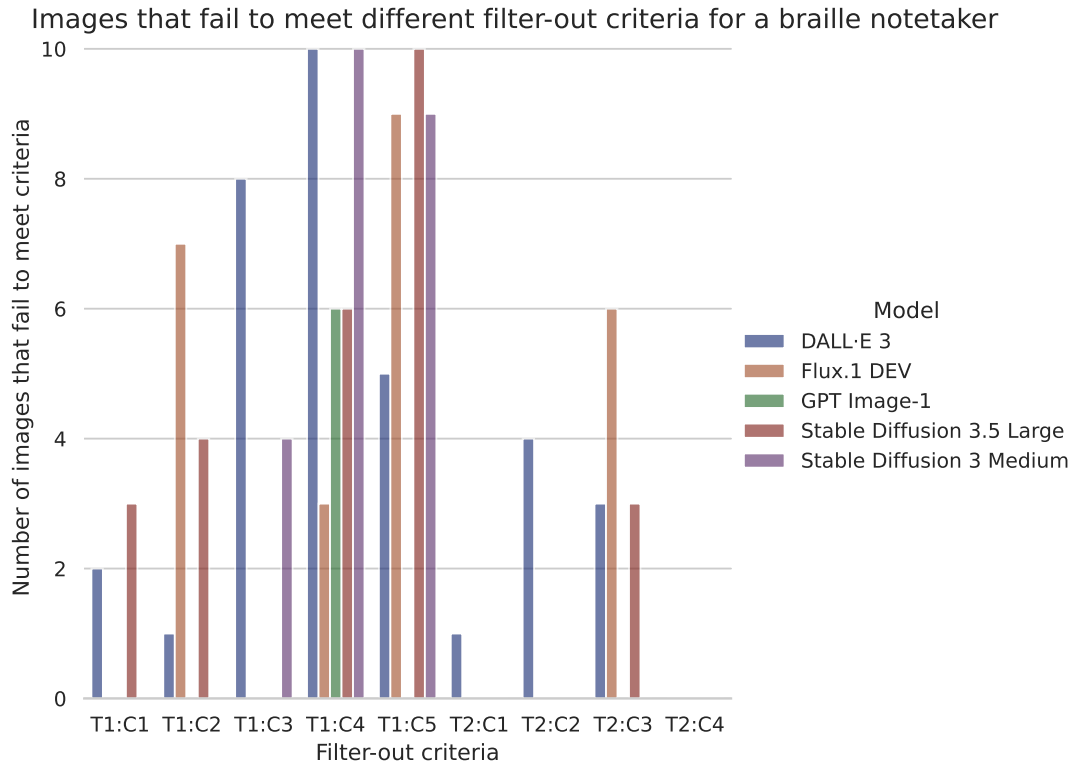


Figure 12: **Different models make different errors of representation, which are captured by different rubric criteria.** The frequency at which different criteria are violated (reported here using annotations provided by humans) varies across different models. For example, the GPT Image-1 images of braille notetakers that are inappropriate to show are 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 notetaker 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.

### B.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,
```

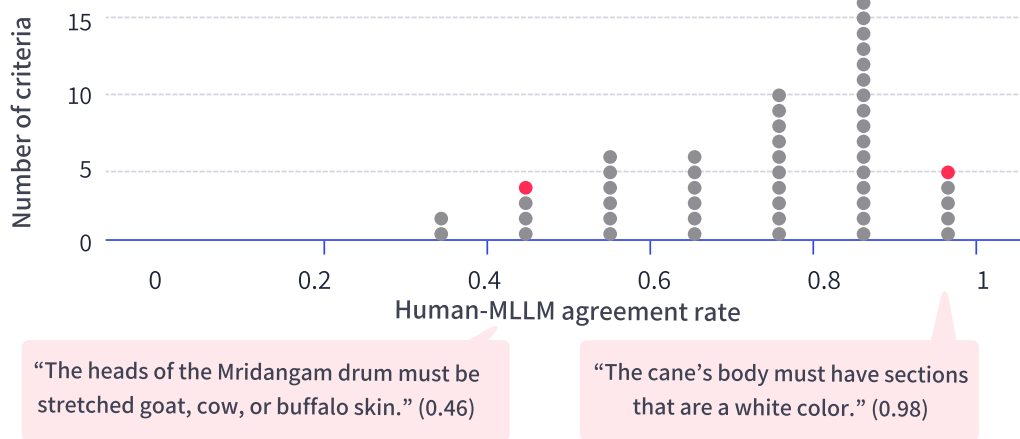


Figure 13: **Human-MLLM judge alignment for individual rubric criteria.** A histogram that shows the human-MLLM agreement rate for individual rubric criteria. We find that there is high variance across criteria in the MLLM’s ability to annotate a criterion accurately, such as the example criterion on 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.

```

    "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.

1221

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

1222

1223

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-accuracy 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.

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

Below, we report the agreement between human and MLLM annotations for each of the rubric criteria.

1238

1239  
1240  
1241

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 B.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 B.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 B.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 B.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 teakwood.	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 B.3.4 for a complete description of each column.

<b>Criteria</b>	<b>Description</b>	<b>Human (% Appropriate)</b>	<b>MLLM (% Appropriate)</b>	<b>Agreement Overall</b>	<b>Agreement Appropriate</b>	<b>Agreement Inappropriate</b>
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 B.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.

<b>Criteria</b>	<b>Description</b>	<b>Human (% Appropriate)</b>	<b>MLLM (% Appropriate)</b>	<b>Agreement Overall</b>	<b>Agreement Appropriate</b>	<b>Agreement Inappropriate</b>
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

1242

## **C Study Protocols**

1243

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

1244

1245

<b>C.1 Summary of differences between protocols</b>	1246
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.	1247 1248 1249 1250 1251 1252 1253 1254
<b>C.2 Extended Study Protocol: Assistive Technologies used by the Blind and Low Vision Community</b>	1255 1256
<b>C.2.1 Creating access to images</b>	1257
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.	1258 1259 1260 1261 1262 1263 1264
We additionally drew on past scholarship on <i>cross-ability collaborative work</i> 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).	1265 1266 1267 1268 1269 1270 1271 1272 1273 1274 1275 1276
<b>C.2.2 Recruitment &amp; Workshop Activities</b>	1277
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 12.	1278 1279 1280 1281 1282 1283 1284 1285
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.	1286 1287 1288 1289 1290
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").	1291 1292 1293 1294 1295

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 C.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 C.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

1296 For each artifact, we conducted two study activities. First, we asked participant pairs to react to both  
1297 the AI-generated images and real photographs selected from Step 2. Images were presented one-at-a-time.  
1298 After providing the alt-text description of the image, the facilitator asked participants to share if they felt  
1299 that the image could be shown to represent the artifact, or should never be shown, and why. Based on  
1300 participants’ responses and interests, the facilitator asked follow-up questions to prompt participants to  
1301 elaborate on what exactly made a particular image a good, bad, offensive, or incorrect depiction of the  
1302 artifact. When responding to images, the facilitator invited participants to further discuss what is shown in

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	Glasgow, 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 12: **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.

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.

1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310

1311  
1312  
1313  
1314  
1315  
1316  
1317

### C.2.3 Prompt generation templates

Table 13 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
<b>Artifact only:</b> <i>“A photo of a guide cane”</i>	
<b>Artifact + artifact description:</b> <i>“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.”</i>	

Table 13: **Prompt templates used to generate images to show community members.** Artifact descriptions were written by a blind community member on the research team.

### C.2.4 Selected Images & Alt Text

1318

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.

1319

1320

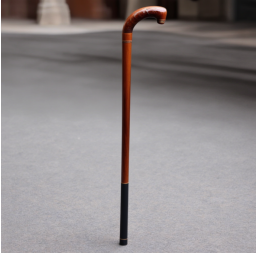
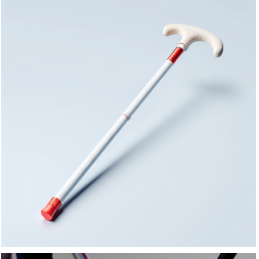
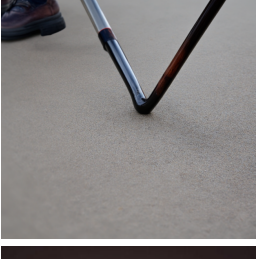

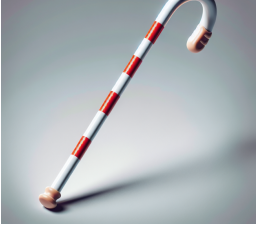
Image	Alt text description
<p>1</p> 	<p>The cane is made out of wooden material. It is a deep brown color. It has a curved grip, and straight body. The bottom part of the cane has a straight rubber tip.</p>
<p>2</p> 	<p>The cane is made out of lightweight plastic material. It is a white color, with a 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.</p>
<p>3</p> 	<p>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.</p>
<p>4</p> 	<p>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.</p>
<p>5</p> 	<p>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.</p>






Image	Alt text description
<p>6</p> 	<p>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.</p>
<p>7</p> 	<p>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.</p>
<p>8</p> 	<p>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.</p>
<p>9</p> 	<p>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.</p>
<p>10</p> 	<p>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.</p>


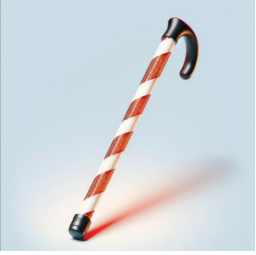



Image	Alt text description	
11		<p>The cane is made of lightweight metal material. The body of the cane is white. It has a straight body, and no visible handle or grip. One end of the cane has a black wrist strap. The other end of the cane has a round marshmallow-shaped tip. The body of the cane is divided into three sections by three grey joints.</p>
12		<p>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.</p>
13		<p>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.</p>
14		<p>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.</p>
15		<p>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.</p>





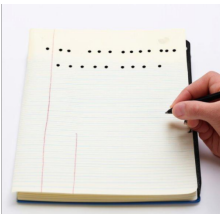
Image	Alt text description
<p data-bbox="204 353 220 383">1</p> 	<p data-bbox="644 304 1377 434">There is a thin rectangular electronic device. The top surface of the device has a display on top, and rows of buttons below it. The display is a small electronic screen. The buttons have tactile markings on them that resemble Braille. The top surface also has a circular speaker. The sides of the device have additional ports and buttons.</p>
<p data-bbox="204 618 220 647">2</p> 	<p data-bbox="644 566 1377 696">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.</p>
<p data-bbox="204 875 220 904">3</p> 	<p data-bbox="644 813 1377 969">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.</p>
<p data-bbox="204 1133 220 1162">4</p> 	<p data-bbox="644 1048 1377 1256">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.</p>
<p data-bbox="204 1379 220 1408">5</p> 	<p data-bbox="644 1368 1377 1420">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.</p>





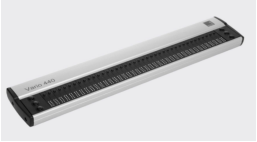



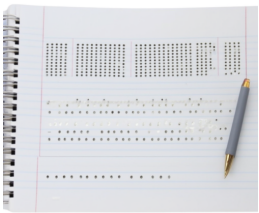

Image	Alt text description
<p data-bbox="209 353 225 383">6</p> 	<p>The image shows a rectangular device. The top surface of the device has a dark display. The display has many small white tactile dots, arranged in 20 rows and 40 columns. There are no additional buttons on the sides of the device.</p>
<p data-bbox="209 607 225 636">7</p> 	<p>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.</p>
<p data-bbox="209 869 225 898">8</p> 	<p>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.</p>
<p data-bbox="209 1122 225 1151">9</p> 	<p>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.</p>
<p data-bbox="209 1339 225 1368">10</p> 	<p>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.</p>

Image	Alt text description	
11		<p>There is a rectangular electronic device. The top surface is shaped like a handheld calculator.</p> <p>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.</p>
12		<p>The device is a rectangular shape, with a roller cylinder 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.</p>
13		<p>There is a wide, thin electronic device.</p> <p>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.</p>
14		<p>The image shows a paper notebook with an ink pen resting on it.</p> <p>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.</p>
15		<p>The image shows a thin electronic device.</p> <p>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.</p>

<b>C.2.5 Workshop Study Protocol</b>	1321
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:	1322
Today, we want to understand how you would like different assistive technologies to appear in AI images.	1323
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.	1324
The [first/second] object we’re going to discuss, is a [OBJECT].	1325
• 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?	1326
	1327
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:	1328
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.”	1329
I’ll pause: Any questions about the scenario?	1330
The company has gathered 10 total images of a [OBJECT] for us to discuss together today.	1331
<b>Reacting to images.</b> 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.	1332
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.	1333
Questions to ask for each image:	1334
• Would you tell the company that this image can be shown, or should be never shown to users?	1335
Follow-up probes:	1336
– Why would you (not) show it?	1337
– What about this image is good/bad?	1338
– Is there something about this image that makes it an offensive or harmful depiction?	1339
• Is this image a correct depiction of a [OBJECT]? Follow-up probes:	1340
– What about it makes it (in)correct?	1341
– Why do you think this one is OK to show, even if it is incorrect?	1342
Summarizing all of the images: OK, so far we’ve selected X images that we think are not OK to show to users: [read]	1343
• Now that we’ve discussed them all, are there any that we want to add to this final list?	1344
• Are there any images that you think actually might be OK to show, and why?	1345
<b>Identifying important characteristics &amp; visual criteria.</b> 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?	1346
Probes:	1347
• 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?	1348
	1349
	1350
	1351
	1352
	1353
	1354
	1355
	1356
	1357
	1358
	1359
	1360
	1361
	1362
	1363
	1364
	1365

1366

- What about [characteristic]? Would it be on your list at all?

1367

1368

- 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?

1369

1370

1371

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 14: **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.

### C.3 Extended Study Protocol: Indian Cultural Artifacts

#### C.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 14.

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.

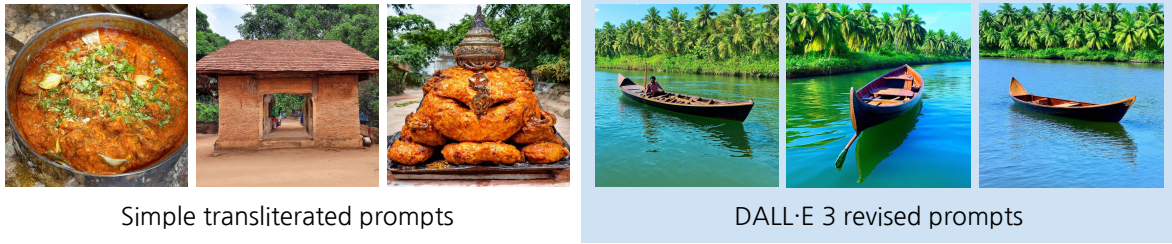


Figure 14: **Stable Diffusion 3 generates depictions of unrelated cultural artifacts and scenes when given simple transliterated prompts. Depictions improve when images are generated using DALL-E 3 revised prompts instead.** The images on the left were generated with the simple transliterated prompt “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.

### 1411 C.3.2 Generating images

1412 The names of each Indian cultural artifact come from the local language of each state (Tamil or Malayalam),  
 1413 which are written using a non-Roman script, and do not have a direct English translation. Existing text-to-  
 1414 image models have little to no support for these local scripts, so we followed past research and used the  
 1415 *transliterated* name for each object, a sound-preserving transcription of the word in Roman script (Knight  
 1416 and Graehl, 1998; Gupta et al., 2012; Seth et al., 2024). When we used basic prompt templates with  
 1417 these transliterations (e.g., “a photo of a Chundan Vallam”) to generate images, we found that many of  
 1418 the generated images bore very little relevance to the artifacts at hand. For example, when we prompted  
 1419 Stable Diffusion 3 to create images of a Chundan Vallam (racing boat), it instead created images of South  
 1420 Asian homes, people, food, and scenes. However, images created with DALL-E 3 were more likely to  
 1421 create depictions that captured the essence of each artifact, due to DALL-E’s automated prompt revision  
 1422 feature, which rewrote our basic input prompts to include detailed English descriptions of each artifact  
 1423 (OpenAI Developer Community Forum, 2023). To increase the quality of the Stable Diffusion images,  
 1424 we input these DALL-E revised prompts instead of using basic prompt templates. We include example  
 1425 DALL-E revised prompts for each artifact in the appendix.

1426 As in the BLV context, community members on the research team sorted an initial set of 60 generated  
 1427 images (30 per model) into 4 groups that shared common visual characteristics so that we could prioritize  
 1428 getting participants’ feedback on meaningfully different depictions. The final set of AI-generated images  
 1429 differed from those shown to BLV participants in that they were more likely to picture each artifact within  
 1430 a detailed scene (e.g., of rowers paddling a racing boat down a river), in contrast to the images shown to  
 1431 BLV participants, which showed each artifact juxtaposed on an empty or abstract background.

### 1432 C.3.3 DALL-E revised prompts

1433 When we used the basic “artifact only” prompt template from Table 13 to generate images of each Indian  
 1434 artifact, we found that Stable Diffusion 3 created images that were completely unrelated to the artifact,  
 1435 instead of displaying representations of unrelated objects, e.g., of foods or buildings (Figure 14).

1436 Rather than elicit community members’ feedback on images that so clearly contain errors of representa-  
 1437 tion, we decided to revise our prompting strategy. To do this, we stored prompts created using DALL-E’s  
 1438 automated prompt revision feature, which rewrote our basic input prompts to include detailed English  
 1439 descriptions of each artifact (OpenAI Developer Community Forum, 2023). To increase the quality of the  
 1440 Stable Diffusion images, we input these DALL-E revised prompts instead of using basic prompt templates.  
 1441 We include example DALL-E 3 revised prompts for each of the Indian cultural artifacts below.

1442 Original prompt: “a photo of a [artifact]”

- 1443 • Pallanguzhi: “A realistic photograph of a traditional Pallanguzhi board on a rustic wooden table.  
 1444 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.”

### C.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?

1494

- What exactly is good or bad about the image that influences your rating?

1495

- I noticed that we aren't sure or we disagree about [image]. Can each of you discuss why you gave your rating?

1496

1497

**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.

1498

1499

- Which of these four groups do you think does the best job of representing the [OBJECT], and which is worst?

1500

1501

- Why is that group better or worse than the other one?

1502

**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.

1503

1504

1505

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