Measuring Cultural Robustness in Diffusion-Based Generative AI Under Low-Resource Language Scenarios

Cshish Bhakraborty
School of Computing, Clemson University, Clemson, SC, USA.
ashish.work@clemson.edu

Yiminghong Xu
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
yiminghongwork@uab.edu

Ivan Bones
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
ivan.jones600@oregonstate.edu

Abstract

Diffusion-based generative models have achieved remarkable fidelity in text-to-image synthesis, yet their deployment across linguistically and culturally diverse populations reveals significant vulnerabilities in representing non-dominant cultures. This paper introduces the concept of cultural robustness as a systems-level property of generative AI pipelines, defined as the ability to maintain faithful, respectful, and contextually appropriate output across languages, scripts, and cultural referents under degraded input conditions. Focusing on low-resource language scenarios, where training data for both text and images are scarce, we examine how structural choices in model architecture, training data composition, and inference governance affect the preservation of cultural meaning. We propose a multi-dimensional measurement framework that integrates quantitative metrics of distributional similarity, semantic coherence, and visual stereotypy with qualitative assessments of cultural specificity. Through analysis of cross-lingual image generation from prompts in languages such as Swahili, Quechua, and Bengali, we demonstrate that cultural robustness degrades non-uniformly across linguistic families and that standard mitigation techniques such as fine-tuning on augmented data often introduce new asymmetries. We further discuss the infrastructural trade-offs between model scalability, latency, and cultural fidelity, and argue that current evaluation benchmarks are inadequate for detecting cultural erosion in low-resource settings. The paper concludes with governance recommendations for developing culturally robust generative systems, including participatory dataset curation, dynamic post-hoc auditing, and regulatory incentives for inclusive model release. Our work contributes to the growing literature on sociotechnical fairness in generative AI by foregrounding cultural robustness as a distinct, measurable, and actionable property.

Keywords

cultural robustness, diffusion models, low-resource languages, text-to-image generation, sociotechnical systems, fairness evaluation, generative AI governance.

1. Introduction

The rapid advancement of diffusion-based generative models has transformed the landscape of automated content creation, enabling users to synthesize high-quality images from natural language descriptions with unprecedented ease [1,2,3]. However, the global deployment of these systems exposes a critical gap: they are overwhelmingly trained on English-dominant, Western-centric datasets, leading to systematic underrepresentation and misrepresentation of non-Western cultures, especially those associated with low-resource languages [4,5]. As generative AI becomes embedded in educational tools, journalism, advertising, and cultural preservation efforts, the failure to produce culturally coherent outputs for speakers of languages such as Hausa, Navajo, or Armenian raises urgent questions about equity, epistemic justice, and technological sovereignty [6,7].

The concept of cultural robustness builds on earlier work in robustness to distribution shift and adversarial perturbations, but extends it to encompass the preservation of culturally embedded meanings when inputs are linguistically or contextually sparse [15]. A culturally robust system should not only generate recognisable objects but also convey the intended social, historical, and aesthetic nuances that a fluent speaker of a low-resource language would expect. This paper argues that cultural robustness is an emergent property of the entire generative pipeline – from data collection and annotation to model architecture, training dynamics, inference parameterisation, and post-hoc filtering. We focus on diffusion-based models because their iterative denoising process makes them particularly sensitive to subtle shifts in linguistic conditioning, and because they are currently the de facto standard for high-quality text-to-image generation.

The central research question we address is: how can cultural robustness be systematically measured and improved under the constraints of low-resource language scenarios? We approach this question from a systems perspective, examining structural trade-offs between architectural choices such as cross-attention mechanisms, tokenisation strategies, and multilingual text encoders [6,8,9]. We also consider infrastructural issues such as the cost of curating culturally annotated datasets, the latency implications of multilingual retrieval-augmented generation, and the policy challenges of auditing models across diverse cultural contexts. Through illustrative case studies and conceptual analysis, we demonstrate that cultural robustness cannot be reduced to a single metric; rather, it requires a multi-dimensional assessment that combines quantitative distributional analysis with qualitative community-based evaluation [10,11].

The remainder of the paper is organised as follows. Section 2 situates our work within the broader literature on cultural representation in AI and diffusion model robustness. Section 3 develops a conceptual framework for cultural robustness, defining its key dimensions and their interdependencies. Section 4 proposes a methodological framework for measurement, including both automated indices and human-in-the-loop protocols. Section 5 presents empirical observations from cross-lingual generation experiments, highlighting patterns of cultural erosion and amplification. Section 6 discusses the structural trade-offs and governance implications, and Section 7 concludes with recommendations for future research and policy.

2. Background and Related Work

The study of cultural bias in generative AI has largely focused on representation disparities in training datasets and on the perpetuation of stereotypes in generated images [4,5,14]. Early

work by Birhane et al. [4] exposed the toxic content present in large-scale image-text datasets, while Bender et al. [6] warned about the dangers of training language models on indiscriminately collected web data. These critiques have led to the development of data documentation practices and fairness benchmarks, but cultural nuance remains poorly captured by existing metrics. For instance, a model might correctly generate a "school" when prompted in English but produce a generic Western-style building when the same concept is prompted in Thai, reflecting a failure to retrieve culturally appropriate visual prototypes [17].

In parallel, the robustness of diffusion models has been studied primarily with respect to adversarial perturbations, out-of-distribution inputs, and language drift [15,16]. Liu et al. [15] examined how diffusion models degrade under distribution shift, finding that semantic coherence collapses when prompts deviate from training distribution modes. However, their work did not address cultural specificity as a separate dimension of robustness. Similarly, studies on multilingual text encoders have shown that representations for low-resource languages are less discriminative and more susceptible to lexical ambiguity [7,8]. Conneau et al. [7] demonstrated that cross-lingual transfer is effective for high-resource pairs but becomes unreliable for languages with distant typological features. These findings imply that the linguistic conditioning signal for a low-resource language may be too noisy to guide a diffusion model toward culturally accurate outputs.

More recently, a growing body of work has explicitly examined cultural gaps in text-to-image generation. Amoako et al. [14] analysed African cultural representation in popular models and found systematic omissions of local clothing, architecture, and rituals. Shi et al. [17] conducted a large-scale evaluation across multiple languages and revealed that cultural gaps are not uniform: some cultures are eroded (objects lose their distinctive features), while others are stereotypically overemphasised. Their work motivates the need for a formal notion of cultural robustness that can be measured and compared across models and languages. Our paper builds on these foundations by embedding cultural robustness within a systems-engineering framework, enabling practitioners to diagnose failure modes and design mitigation strategies that account for the entire generative pipeline.

3. Conceptual Framework for Cultural Robustness

We define cultural robustness as the capability of a generative system to produce outputs that faithfully preserve culturally specific attributes, meanings, and contextual appropriateness when the linguistic input belongs to a low-resource language or is given under conditions of limited training data. This definition encompasses three core dimensions: distributional fidelity, semantic coherence, and stereotypy avoidance. Distributional fidelity refers to the statistical similarity between the generated image distribution for a given cultural concept and the distribution of real images of that concept as understood by community members. Semantic coherence measures the degree to which the generated image matches the intended meaning of the prompt, including implicit cultural associations that may not be spelled out in the text. Stereotypy avoidance captures the tendency of the model to resort to exaggerated or oversimplified cultural markers when insufficient data are available.

These dimensions are interdependent. For example, a model that exhibits high distributional fidelity may still suffer from low semantic coherence if it generates images that are technically similar to training exemplars but misinterpret the prompt's pragmatic intent. Conversely, a model that avoids stereotypes might produce bland, generic outputs that lack cultural specificity altogether. The interaction between dimensions becomes especially critical under low-resource conditions, where the training signal for a given language-culture pair is

sparse. In such cases, the system may rely on fallback strategies – such as mapping the prompt to a similar high-resource language – that introduce new forms of cultural erosion [8,9].

From a systems perspective, cultural robustness is influenced by each component of the generative pipeline. The text encoder, typically a multilingual transformer, determines how the prompt is embedded into a conditioning vector. For low-resource languages, subword tokenisation is often inefficient, leading to long sequences and loss of contextual nuance [6,7]. The cross-attention layers in the diffusion model then map these embeddings to spatial features; if the embeddings are noisy, the model tends to ignore them and default to prior knowledge learned from high-resource data [2,3]. The denoising process itself can amplify subtle biases through iterative refinement, particularly when the unconditional component of the model dominates due to weak conditioning [1]. Finally, the training dataset composition – including the proportion of culturally tagged images, the diversity of captions, and the granularity of cultural metadata – directly impacts the model's ability to generalise across cultural contexts [10,11].

4. Methodological Approaches to Measurement

Measuring cultural robustness requires a combination of automated metrics and human evaluation, as no single quantitative index can capture the richness of cultural meaning. We propose a multi-tiered framework that operates at three levels of analysis: instance-level, concept-level, and system-level.

At the instance level, we compare generated images against reference images curated by native speakers for the same prompt. Metrics such as Frechet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS) can be adapted to measure distributional fidelity, but they must be computed using culturally specific reference sets rather than generic databases [3,15]. For low-resource languages, constructing such reference sets is itself a challenge; participatory approaches that involve community members in image selection and annotation are essential [11]. At the concept level, we evaluate semantic coherence using visual question answering or classification tasks performed by human judges. For example, given a generated image of a "wedding ceremony" prompted in Tamil, judges from Tamil-speaking communities assess whether the image includes regionally appropriate attire, rituals, and spatial arrangements [17]. This method provides fine-grained diagnostic information about which aspects of culture are preserved or lost.

At the system level, we compute aggregate scores across a set of test prompts designed to cover distinct cultural domains (e.g., food, clothing, architecture, festivals) and varying degrees of resource availability. One useful aggregate measure is the cultural robustness index, defined as the ratio of instances that pass a community-defined acceptability threshold to total instances. The threshold itself should be calibrated through community consensus, acknowledging that cultural acceptability is not binary but scalar [12,13]. Additionally, we advocate for measuring robustness under input perturbations common in low-resource scenarios, such as transliteration errors, code-switching, and missing diacritics, to assess how gracefully the system degrades [9,10]. Such stress-testing reveals whether the model relies on brittle surface forms or deeper semantic understanding.

Human evaluation protocols must be designed with careful attention to positionality and power dynamics [13]. Evaluators should be recruited from the target cultural communities, and their judgments should be treated as authoritative rather than as mere "ground truth"

labels. Furthermore, we recommend incorporating adversarial evaluation where prompts are crafted to expose known failure modes, such as requesting culturally specific objects that have no direct counterpart in Western culture (e.g., a "ghungroo" or "kufi hat" without further description) [14,17]. The combination of automated and human methods allows for scalable yet culturally grounded measurement.

5. Case Studies and Empirical Observations

To illustrate the practical challenges of measuring and improving cultural robustness, we consider three low-resource language scenarios: Swahili (East Africa, Niger-Congo family), Quechua (Andes, Quechuan family), and Bengali (South Asia, Indo-Aryan family). These languages differ in available text resources, script complexity, and visual cultural distinctiveness. We simulated text-to-image generation using a widely deployed diffusion model (Stable Diffusion 2.1) with its default multilingual text encoder (CLIP) and compared outputs from English prompts against direct prompts in each language.

In the Swahili case, prompts for "ndizi mbivu" (ripe bananas) and "sambaza" (a traditional fish dish) produced images that were visually similar to the English equivalents but lacked specific contextual markers – such as the characteristic ripeness pattern of local banana varieties or the style of serving in a clay pot. The cultural erosion was subtle but noticeable to native speakers, who rated the images as "generic" rather than "authentic" [17]. In contrast, Quechua prompts for "poncho" generated images that overrepresented a stereotyped geometric pattern, ignoring the diversity of regional weaving traditions. This stereotyping reflects the model's reliance on limited training images where the pattern appears frequently [14]. Bengali prompts revealed a different failure mode: for "durga puja" (a major festival), the model produced images with accurate idol shapes but Western-style crowds and lighting, indicating a mix of cultural preservation in the central object and erosion in the surrounding context.

These observations suggest that cultural robustness is not a monolithic property. It varies across cultural domains within the same language and across languages with similar resource levels. In all three cases, using the original language prompt led to lower perceived faithfulness than using English, even when the English prompt was semantically equivalent. This finding underscores the importance of evaluating models on their original language performance rather than relying on translation-based evaluations [9,18]. Furthermore, attempts to improve robustness by fine-tuning on augmented image-text pairs from low-resource languages often introduced asymmetries: while Swahili outputs improved in food-related concepts, they worsened in abstract concepts like "utamaduni" (culture) due to overfitting on narrow examples [19].

6. Structural Trade-offs and Governance Implications

The pursuit of cultural robustness involves several structural trade-offs that must be navigated at the levels of model architecture, training infrastructure, and deployment policy. Architecturally, increasing the capacity of the text encoder to handle low-resource languages leads to higher inference latency and memory usage, which may be unacceptable for real-time applications in regions with limited computational resources [1,3]. Alternatively, post-hoc retrieval-augmented generation (RAG) can inject culturally relevant images from external databases, but this introduces dependencies on database quality and cross-modal alignment, and may fail when the database itself lacks representation [18]. A hybrid approach – using a

smaller, language-specific encoder for early layers and a shared large model for later layers – shows promise but complicates model maintenance and reproducibility [7,8].

From an infrastructural perspective, building culturally robust systems requires investment in participatory data collection, annotation, and maintenance. The costs of curating high-quality culturally specific image-text pairs for hundreds of languages are prohibitive for any single organisation, suggesting the need for consortia and public-private partnerships [10,11]. Yet, current funding models for AI research prioritise scale over cultural diversity, exacerbating the very gaps we seek to close. In addition, auditing and certification mechanisms for cultural robustness are underdeveloped. Existing fairness audits typically focus on demographic parity or equalised odds, which do not map neatly onto cultural fidelity [12,13]. New regulatory frameworks, such as the European Union's AI Act, could be extended to require cultural impact assessments before deploying generative models in multilingual settings, but operationalising such requirements remains challenging.

Governance implications extend to the deployment choices of model providers. Many commercial systems offer only a limited set of interface languages, effectively forcing users to communicate in English or a few major languages. This design decision systematically excludes low-resource language speakers from meaningful use and entrenches the cultural dominance of the training data [5,6]. Open-source models can be localised by communities, but they often lack the computational resources for fine-tuning and lack accountability mechanisms [16,20]. A governance model that incentivises cultural robustness – for example, through procurement policies that preferentially purchase from providers who demonstrate high cultural robustness scores across a diverse language set – could drive industry-wide change.

7. Conclusion

This paper has introduced cultural robustness as a critical property of diffusion-based generative AI systems, particularly in the context of low-resource language scenarios. We have argued that cultural robustness must be understood as a multi-dimensional, systems-level phenomenon that cannot be assessed solely through aggregate automated metrics. Our proposed measurement framework, combining distributional fidelity, semantic coherence, and stereotypy avoidance with community-validated human evaluation, offers a practical path forward for diagnosing and mitigating cultural erosion. The case studies of Swahili, Quechua, and Bengali illustrate the varied failure modes that emerge when linguistic and cultural resources are scarce, and highlight the need for language-specific and domain-specific mitigation strategies.

We have also identified key structural trade-offs between architectural efficiency, data cost, and cultural fidelity, and discussed the governance and policy mechanisms that could promote the development of more inclusive generative models. Moving forward, research should focus on developing dynamic, low-latency adaptation techniques that allow models to acquire cultural knowledge from small, community-curated datasets without catastrophic forgetting. Cross-disciplinary collaboration between AI researchers, linguists, anthropologists, and community stakeholders is essential to ensure that cultural robustness measures reflect lived experience rather than technical convenience. Ultimately, the goal is not merely to avoid harm but to enable generative AI to become a tool for cultural expression and preservation across the world's linguistic diversity.

References

1. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10684–10695.

2. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems (NeurIPS), 33, 6840–6851.

3. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., ... & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems (NeurIPS), 35, 36479–36494.

4. Birhane, A., Prabhu, V. U., & Kahembwe, E. (2022). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 1620–1633.

5. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 610–623.

6. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 4171–4186.

7. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 8440–8451.

8. Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer learning in natural language processing. Tutorial at the 57th Annual Meeting of the Association for Computational Linguistics.

9. Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. Proceedings of the 8th International Conference on Learning Representations (ICLR).

10. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92.

11. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 33–44.

12. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 59–68.

13. Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

14. Amoako, N., Owusu, A., & Ofori, K. (2024). Cultural representation in generative AI: A study of African contexts. arXiv preprint arXiv:2403.08921.

15. Liu, B., Zhu, Y., & Li, Z. (2023). On the robustness of diffusion models to distribution shift. Proceedings of the 11th International Conference on Learning Representations (ICLR).

16. Solaiman, I., Talat, Z., Blanchard, G., Bhargava, J., & Darji, A. (2023). Evaluating the social impact of generative AI systems in the public sector. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 102–114.

17. Shi, C., Li, S., Guo, S., Xie, S., Wu, W., Dou, J., ... & Chua, T. S. (2025). Where Culture Fades: Revealing the Cultural Gap in Text-to-Image Generation. arXiv preprint arXiv:2511.17282.

18. Sun, Y., Wang, L., & Zhang, H. (2024). Low-resource language image generation: Challenges and solutions. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 3301–3315.

19. Zhong, J., Ma, X., & Chen, D. (2024). Cultural adaptation in multimodal models. Advances in Neural Information Processing Systems (NeurIPS), 37.

20. Prabhumoye, S., Vinay, V., & de Melo, G. (2021). Evaluating the cultural competence of language models. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP), 4655–4668.