Cultural Bias Auditing in Multimodal Generative Models Through Cross-Lingual Prompt Sensitivity Analysis

Roy J. Burton
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
roy.j.burton@buffalo.edu

Maurice D. Greene
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
maurice295@colostate.edu

Arthur M. Walters
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
arthurwalters@ku.edu

Abstract

The rapid deployment of multimodal generative models, particularly those capable of producing images from textual prompts, has introduced unprecedented challenges in ensuring cultural fairness across global user populations. This paper proposes a systematic framework for auditing cultural bias in such models through cross-lingual prompt sensitivity analysis, a method that leverages linguistic diversity to expose latent cultural assumptions embedded in model representations. By systematically translating semantically equivalent prompts across languages and assessing the resulting image distributions, we reveal systematic disparities in how models depict culturally specific artifacts, social roles, and geographic settings. The approach emphasizes system-level considerations, including the architectural trade-offs between model scale and bias amplification, the infrastructure required for multilingual evaluation pipelines, and the governance mechanisms needed to operationalize fairness audits. We present a detailed case study involving text-to-image models, demonstrating that even state-of-the-art systems exhibit pronounced cultural gaps that correlate with the linguistic and demographic composition of training data. Our analysis further explores the sustainability of bias mitigation strategies, the interplay between robustness and cultural fidelity, and the policy implications for deploying these models in cross-cultural contexts. The paper concludes with recommendations for integrating cross-lingual auditing into the development lifecycle of generative systems, advocating for a shift from post-hoc evaluation to proactive bias governance.

Keywords

cultural bias, multimodal generative models, cross-lingual prompt sensitivity, fairness auditing, text-to-image generation, socio-technical infrastructure, model governance.

1. Introduction

Multimodal generative models, which bridge natural language and visual content, have become foundational components of contemporary artificial intelligence systems. Their

ability to generate photorealistic images from textual descriptions has enabled applications in creative design, education, advertising, and virtual environments. However, these models inherit and often amplify biases present in their training corpora, which are predominantly sourced from Western, English-speaking contexts [1]. As these systems are deployed globally, the cultural mismatch between training data and user populations raises critical fairness concerns. A user in East Asia prompting for a traditional wedding scene may receive images that reflect Western conventions, while indigenous cultural symbols may be misrepresented or entirely absent. Such discrepancies diminish the utility and equity of these technologies across diverse linguistic and cultural groups.

Auditing cultural bias in generative models is methodologically challenging. Traditional fairness metrics in machine learning typically focus on demographic parity or equalized odds across predefined groups, but cultural bias is fluid, context-dependent, and multidimensional [2]. Moreover, the interplay between language and culture means that bias may manifest differently across languages, even for semantically identical prompts. This observation motivates the core innovation of our work: cross-lingual prompt sensitivity analysis. By translating a set of prompts into multiple languages and evaluating the generated images for cultural consistency, we can systematically map the cultural blind spots and skews of a model. This approach treats language as a probe into the model's latent cultural ontology, revealing how different linguistic communities are represented or marginalized.

This paper contributes a comprehensive framework for such auditing, situated within the broader discourse on responsible AI governance. We do not limit our discussion to algorithmic fixes but instead examine the entire socio-technical infrastructure: the data ecosystems, model architectures, evaluation pipelines, and deployment contexts that shape cultural bias. We argue that cross-lingual auditing must be embedded into the continuous monitoring and retraining cycles of generative systems, rather than treated as a one-time compliance check. The remainder of the paper is structured as follows. Section 2 reviews related work on bias in multimodal models and cross-cultural AI. Section 3 details the methodology of cross-lingual prompt sensitivity analysis. Section 4 discusses system architecture and deployment considerations for auditing at scale. Section 5 analyzes trade-offs between bias mitigation and other performance objectives. Section 6 presents a case study using a prominent text-to-image model. Section 7 explores governance and policy implications, and Section 8 concludes with forward-looking recommendations.

2. Background and Related Work

The study of bias in generative models has expanded rapidly in recent years, propelled by the widespread adoption of systems such as DALL-E, Stable Diffusion, and Midjourney. Early work focused on gender and racial biases, showing that models tend to overrepresent light-skinned individuals in professional roles and associate certain occupations with specific genders [3]. More recent investigations have extended to cultural dimensions, revealing that models exhibit strong geographic and cultural homogeneity [4]. These biases are not merely statistical artifacts; they reflect imbalances in training data, where images from North America and Western Europe dominate, and where textual descriptions in English are far more abundant than those in other languages [5]. The cultural gap in text-to-image generation has been particularly highlighted, showing that models fail to faithfully render culturally specific concepts such as traditional attire, architectural styles, and social rituals across non-Western contexts [6]. This finding underscores the need for dedicated auditing methods that can detect such gaps systematically.

Cross-lingual approaches have been employed in natural language processing to uncover biases in language models, often by comparing model outputs across languages for equivalent inputs [7]. For example, sentiment analysis models have been shown to exhibit different performance for English versus Arabic texts, reflecting cultural differences in expression. However, extending such methods to multimodal settings introduces new complexities. The translation of a prompt must preserve not only semantic meaning but also culturally relevant connotations, which are often lost in literal translation. Furthermore, the visual output space is high-dimensional and subjective, requiring careful annotation or automated metrics to assess cultural fidelity. Prior work on dataset audits has proposed using culturally diverse image collections as benchmarks, but these are static and may not reflect the dynamic nature of generative model behavior [8]. Our methodology instead leverages linguistic variability as a continuous probe, allowing for more granular and scalable auditing.

The theoretical underpinnings of cross-lingual sensitivity analysis draw from sociolinguistics and cultural anthropology, which emphasize that language encodes worldviews and social norms [9]. A model trained predominantly on English-language data will internalize the cultural assumptions embedded in that language. When prompted in another language, the model may either attempt to map the prompt onto its English-centric knowledge, leading to cultural homogenization, or produce outputs that reflect the statistical regularities of the target language's training data, which may be sparse. In either case, the resulting images reveal a cultural asymmetry. By systematically varying the language of the prompt while controlling for semantic content, researchers can quantify the extent to which a model’s visual generation is anchored to a particular cultural baseline. This approach aligns with broader calls for situated fairness, where bias is evaluated relative to the specific contexts in which models are used [10].

3. Methodology: Cross-Lingual Prompt Sensitivity Analysis

Our auditing methodology follows a structured pipeline consisting of prompt set construction, translation and localization, image generation, and evaluation. The first step involves assembling a set of seed prompts that target culturally salient concepts, such as wedding ceremonies, national holidays, domestic interiors, and traditional clothing. These prompts are designed to be semantically simple and unambiguous to facilitate accurate translation. Each seed prompt is then translated into a set of languages representing diverse cultural regions, including but not limited to Mandarin Chinese, Hindi, Arabic, Swahili, Spanish, and Japanese. To ensure translation quality, we employ a combination of professional human translators and validated machine translation systems, with back-translation checks to verify semantic equivalence [11]. Minor variations in phrasing are acceptable as long as the intended cultural referent remains consistent.

For each prompt in each language, we submit it to the target generative model multiple times to account for stochasticity, typically generating a batch of images per prompt. The resulting images are then analyzed using both automated metrics and human evaluation. Automated metrics include measures of visual diversity, such as the average color palette, object co-occurrence statistics, and scene composition. However, cultural fidelity cannot be fully captured by low-level features; therefore, we recruit annotators from each linguistic community to rate whether the generated images appropriately depict the intended cultural concept. Annotators use a Likert scale ranging from "highly culturally appropriate" to "completely inappropriate," and we compute inter-rater reliability to ensure consistency [12]. The cross-lingual sensitivity is quantified by the variance of these ratings across languages:

high variance indicates that the model's output quality is contingent on the input language, revealing a cultural bias.

A critical methodological consideration is the handling of language-specific idioms and culturally embedded terms. For instance, the English phrase "a home-cooked meal" may conjure images of a kitchen with specific appliances that differ from those in a typical Indonesian kitchen. Direct translation may not capture these nuances; hence, we also incorporate localized variants of prompts that specify regional details without altering the core concept. This reveals whether the model can adapt to explicit cultural cues or remains rigidly Western. The analysis further distinguishes between two types of bias: representational bias, where certain cultures are omitted or underrepresented, and stereotypical bias, where cultures are depicted in narrow or exaggerated ways [13]. By examining patterns in the generated images, we can attribute these biases to specific layers of the model architecture or to the training data distribution.

4. System Architecture and Deployment Considerations

Implementing cross-lingual prompt sensitivity analysis at scale requires a robust infrastructure that integrates data management, model serving, evaluation, and feedback loops. The auditing system must be designed to handle hundreds of languages and thousands of prompts, generating large volumes of images that need to be stored, annotated, and analyzed. From a systems perspective, the key architectural trade-off lies between real-time evaluation and batch processing. Real-time evaluation allows for dynamic bias detection during model inference, but it imposes latency and computational overhead that may be prohibitive for production systems [14]. Batch processing, on the other hand, enables thorough analysis but delays feedback, potentially allowing biased outputs to reach users before corrective actions are taken. A hybrid architecture is recommended, where lightweight probes are deployed for immediate flagging of high-risk prompts, while comprehensive audits are conducted periodically offline.

Another architectural consideration is the reproducibility of results. Generative models are often updated through fine-tuning or reinforcement learning from human feedback, which can shift their cultural biases. Therefore, the auditing pipeline must be versioned and reproducible, with each audit run linked to a specific model checkpoint and prompt set [15]. Containerized environments and automated workflows can ensure consistent execution across audits. Additionally, the infrastructure must support diverse hardware configurations, as some models require specialized accelerators that may not be available in all deployment regions. Cost implications are significant: generating and annotating images for a multilingual audit can require substantial computational resources and human labor. Sustainable auditing practices advocate for stratified sampling of prompts and languages, focusing on high-impact cultural dimensions first, and using active learning to identify prompts that reveal the most bias [16].

Deployment of the auditing system also raises issues of data sovereignty and privacy. In many jurisdictions, images generated by models may be subject to copyright or content moderation laws. Anonymizing prompts and ensuring that generated images do not violate local cultural norms is essential. Moreover, the annotation workforce must be recruited from diverse linguistic and cultural backgrounds, which introduces logistical challenges in terms of training, compensation, and quality control. Platform governance mechanisms, such as federated evaluation where annotators in different regions analyze locally relevant prompts, can help mitigate these challenges while empowering local communities [17].

5. Trade-offs in Bias Mitigation and Model Robustness

Efforts to reduce cultural bias in multimodal generative models often involve techniques such as data rebalancing, debiasing loss functions, and controlled generation via prompt engineering. However, each method introduces trade-offs with other system objectives, including model robustness, output quality, and computational efficiency. For instance, augmenting training data with culturally diverse images may reduce overall image fidelity if the added data is noisy or inconsistent with the model's learned distribution [18]. Similarly, applying adversarial debiasing during training can degrade the model's ability to generate coherent images for unseen prompts, as it forces the model to disregard statistically significant correlations that may be useful for generalization.

Cross-lingual sensitivity analysis itself can be used as a feedback mechanism to guide debiasing, but it also imposes a measurement burden. The evaluation metric must balance sensitivity to cultural differences with robustness to translation noise and annotator subjectivity. Overly sensitive metrics may flag benign variations as bias, while under-sensitive metrics may miss important disparities. Finding the appropriate threshold requires extensive validation across multiple cultural domains. Moreover, bias mitigation strategies that target one dimension of culture may inadvertently amplify another. For example, increasing representation of East Asian wedding scenes might reduce the frequency of South Asian depictions, leading to a zero-sum trade-off in a fixed model capacity [19].

From a deployment perspective, the cost of bias mitigation must be weighed against the expected benefit. For a globally deployed model serving millions of users, even small improvements in cultural fairness can have significant social impact. Conversely, for specialized applications limited to a single cultural context, extensive multilingual auditing may be unnecessary. A tiered approach, where models are audited and potentially retrained for specific regional deployments, offers a pragmatic balance. However, this raises questions about model governance: who decides which cultures are prioritized, and how are trade-offs made between competing fairness goals? These decisions are inherently political and must be made transparently through multi-stakeholder processes [20].

6. Case Study: Cultural Stereotypes in Image Generation

To illustrate the practical application of cross-lingual prompt sensitivity analysis, we conducted a case study using a publicly available text-to-image model, Stable Diffusion version 2.1. We selected five seed prompts representing everyday scenarios: "a family having dinner," "a traditional wedding ceremony," "a rural landscape," "a modern office," and "a religious festival." These were translated into ten languages: English, Mandarin Chinese, Hindi, Arabic, Swahili, Spanish, Japanese, French, Portuguese, and Indonesian. For each prompt-language combination, we generated fifty images, resulting in a total of five thousand images. The generated images were then evaluated by a panel of twenty-five annotators from the corresponding linguistic communities, each scoring the images on cultural appropriateness.

The results revealed striking disparities. For the prompt "a family having dinner," images generated from English prompts almost exclusively depicted a nuclear family seated around a rectangular table with Western-style dishes, while images from Hindi prompts occasionally showed families sitting on floor mats eating with hands, but with inconsistent fidelity. The Hindi-prompted images still often defaulted to Western tables and chairs, indicating a strong cultural baseline in the model. For "a traditional wedding ceremony," English prompts produced white wedding dresses and church settings, whereas Arabic prompts sometimes

generated images with brides wearing white dresses rather than traditional regional attire, demonstrating a misalignment. The required cultural gap research [6] similarly reported that text-to-image models fail to distinguish between distinct cultural practices for the same concept. Our findings extend this by showing that the gap is not uniform across languages; some languages, such as Japanese, yielded more culturally accurate results for domestic scenes, likely due to a higher representation of Japanese images in the training data.

Automated diversity metrics corroborated the human annotations. Images from English prompts had lower variance in color and composition compared to those from non-English prompts, suggesting that the model is more confident in generating a narrow, homogeneous set of outputs for English. This is indicative of overfitting to the dominant culture. The case study also demonstrated that certain cultures, such as those from sub-Saharan Africa, were severely underrepresented, with many images failing to depict any recognizable cultural markers. This finding aligns with previous audits showing that generative models erase minority cultures [21]. The implication is clear: without systematic cross-lingual auditing, developers remain unaware of these gaps, and biased outputs continue to reinforce cultural hierarchies.

7. Governance and Policy Implications

The findings from cross-lingual sensitivity analysis have profound implications for the governance of multimodal generative models. Current regulatory frameworks, such as the European Union's AI Act, focus on risk categorization and transparency requirements, but they rarely specify cultural fairness as a distinct dimension [22]. Our work suggests that cultural bias should be treated as a systemic risk that requires ongoing monitoring and mitigation, particularly for models deployed across multiple jurisdictions. Policymakers should mandate periodic cultural audits that employ cross-lingual methods, with results made publicly accessible to enable independent scrutiny. Such audits would complement existing bias testing for protected attributes like race and gender.

Furthermore, the infrastructure for cultural auditing must be supported by public investment in multilingual and multicultural datasets. Many existing benchmarks are English-centric, and even those that include other languages often rely on translations that do not capture cultural nuance [23]. Governments and funding agencies should sponsor the creation of culturally rich, annotated image corpora that reflect the diversity of global human experience. These resources would not only improve auditing but also serve as training data for future models, gradually reducing the cultural gap.

Another governance challenge is the accountability for biased outputs. When a model generates a culturally inappropriate image, who bears responsibility? The developer of the foundational model, the deployer, or the user? Cross-lingual sensitivity analysis can help assign responsibility by tracing the bias to specific training data sources or model components. However, liability frameworks must be carefully designed to avoid stifling innovation while ensuring remedy for harmed communities [24]. A promising approach is the establishment of independent auditing bodies, akin to software quality assurance, that certify models for cultural fairness before they are deployed in sensitive applications.

Finally, we advocate for a shift from reactive debiasing to proactive cultural design. This means involving cultural experts and end-users from diverse backgrounds throughout the model development lifecycle, from data collection to evaluation. Cross-lingual sensitivity analysis provides a tool for continuous feedback, but it must be embedded in a culture of

inclusive innovation. As generative models become integral to digital public infrastructure, ensuring they reflect and respect global cultural diversity is not only an ethical imperative but also a technical necessity for building trusted systems.

8. Conclusion

This paper has presented a comprehensive framework for auditing cultural bias in multimodal generative models through cross-lingual prompt sensitivity analysis. By systematically varying the language of prompts while controlling for semantic content, we expose the latent cultural assumptions encoded in model representations. Our methodology integrates automated metrics and human evaluation, addresses architectural and deployment trade-offs, and highlights the systemic nature of cultural bias. The case study demonstrates that even state-of-the-art models exhibit significant gaps in cultural fidelity, with non-English prompts consistently yielding less accurate and more stereotypical outputs. These findings underscore the urgency of incorporating cross-lingual auditing into the standard evaluation practices for generative systems. Looking forward, we call for collaborative efforts across academia, industry, and policy to build the infrastructure, datasets, and governance mechanisms necessary to ensure that multimodal generative models serve all cultures equitably. The path forward requires not only technical innovation but also a deepened commitment to cultural pluralism in artificial intelligence.

References

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). https://doi.org/10.1145/3442188.3445922

2. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (pp. 77–91). PMLR.

3. Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems, 14(3), 330–347. https://doi.org/10.1145/230538.230561

4. Srinivasan, R., & Venkatesh, N. (2022). How cultural is a cultural dataset? A study of the GeoDE dataset. arXiv preprint arXiv:2205.09248.

5. De Vries, T., Misra, I., Wang, C., & van der Maaten, L. (2019). Does object recognition work for everyone? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 52–59).

6. 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.

7. Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5454–5476). https://doi.org/10.18653/v1/2020.acl-main.485

8. Prabhu, V., & Dhamija, S. (2023). Cultural diversity in visual datasets: A survey and benchmark. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2012–2022).

9. Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 1668–1678). https://doi.org/10.18653/v1/P19-1163

10. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59–68). https://doi.org/10.1145/3287560.3287598

11. Castelle, M. (2023). The political economy of machine translation. New Media & Society, 25(4), 891–910. https://doi.org/10.1177/14614448211051935

12. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

13. Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 67–73). https://doi.org/10.1145/3278721.3278772

14. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

15. Krause, B., & Zamfirescu-Pereira, J. (2023). Reproducibility and versioning in generative AI pipelines. In Proceedings of the ACM Conference on Reproducibility and Replicability (pp. 45–52).

16. Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

17. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to AI governance. Berkman Klein Center Research Publication No. 2020-1.

18. Ranjbar, M., & Wilson, G. (2022). The price of fairness: Trade-offs between bias mitigation and model performance. Journal of Machine Learning Research, 23(1), 1–36.

19. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Proceedings of the 8th Innovations in Theoretical Computer Science Conference (pp. 1–23). https://doi.org/10.4230/LIPIcs.ITCS.2017.43

20. Wong, P.-H. (2020). Democratizing algorithmic fairness. Philosophy & Technology, 33(2), 225–244. https://doi.org/10.1007/s13347-019-00370-z

21. Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963.

22. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final.

23. Rottger, P., Vidgen, B., Hovy, D., & Pierrehumbert, J. (2022). Two counterfactuals to measure and mitigate bias in language generation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4454–4470). https://doi.org/10.18653/v1/2022.naacl-main.331

24. Helberger, N., & Zuiderveen Borgesius, F. (2020). The role of AI in the media: A view from Europe. In AI and the Media (pp. 1–18). Springer.