Adversarial Prompt Robustness and Cultural Representation Consistency in Large-Scale Text-to-Image Systems

Rajesh Galik
Department of Computer Science, University of Houston, Houston, TX, USA.
malik1980@uh.edu

Daominik Rolfe
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
wolfe888@ucf.edu

Suresh A. Roy
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
suresh.a.roy@unr.edu

Abstract

Large-scale text-to-image generative systems have rapidly evolved into sociotechnical infrastructures that shape visual culture and public imagination at global scale. This paper examines two interconnected yet underexplored systemic properties of these platforms: adversarial prompt robustness and cultural representation consistency. Adversarial prompt robustness refers to the capacity of a system to withstand maliciously crafted textual inputs designed to circumvent safety filters, induce toxic imagery, or exploit model biases. Cultural representation consistency denotes the system's ability to produce outputs that equitably and accurately depict diverse cultural contexts, traditions, and identities without systematically erasing certain groups or reinforcing stereotypical portrayals. We analyze these properties not as isolated technical fixes but as emergent outcomes of complex trade-offs across model architecture, training data curation, alignment pipelines, moderation infrastructure, and governance frameworks. The discussion foregrounds how central architectural choices, including contrastive language-image pretraining, latent diffusion, and classifier-free guidance, create latent pathways that simultaneously affect robustness and representation. We argue that adversarial prompt vulnerability and cultural erasure are deeply entangled with the data supply chain and the incentive structures of large-scale deployment. Furthermore, we probe the policy implications of treating these properties as continuous monitoring problems rather than binary compliance checkpoints. The paper advocates for a polycentric governance model that integrates technical red-teaming, cultural auditing, and participatory feedback mechanisms, while remaining attentive to the energy and economic costs of remediation strategies. By reframing prompt robustness and cultural consistency as dual requirements of trustworthy generative infrastructure, the analysis offers a structured lens for researchers, platform architects, and regulators to navigate the next phase of responsible text-to-image system deployment.

Keywords

text-to-image generation, adversarial robustness, cultural representation, fairness, diffusion models, sociotechnical systems, alignment, infrastructure governance.

1. Introduction

The rapid proliferation of large-scale text-to-image models has redefined the boundaries of synthetic media production, embedding these systems within the fabric of digital communication, advertising, education, and artistic expression. Contemporary architectures such as those built upon contrastive language-image pretraining exemplify the capacity to map linguistic descriptions directly into high-fidelity visual outputs [1]. The open release of latent diffusion models further accelerated adoption, turning what was once a niche research capability into a globally accessible creative tool [2]. As these systems transition from experimental prototypes to foundational platform components, the imperative to understand their systemic vulnerabilities and representational biases has become urgent. This paper addresses two critical dimensions that, while often examined separately, jointly determine the trustworthiness of large-scale text-to-image deployments: adversarial prompt robustness and cultural representation consistency. Adversarial prompt robustness captures the ability of a model to resist inputs intentionally crafted to bypass safety guardrails, generate harmful content, or trigger unintended model behaviors. Cultural representation consistency, by contrast, refers to the model's tendency to produce outputs that reflect the breadth of human cultural experience without homogenization, stereotyping, or erasing particular communities. Both dimensions arise from the same underlying sociotechnical substrate, namely the interplay of massive web-scraped training corpora, representation learning objectives, and the layered post-hoc safety mechanisms that characterize production systems.

Foundation models more broadly exhibit emergent behaviors that are difficult to anticipate solely from static benchmarks, a phenomenon well documented in the language domain [3]. The visual generative counterparts inherit many of these challenges, compounded by the multimodality of their conditioning signals and the polysemy inherent in images. The dangers of uncurated internet-scale data have been articulated through critiques of stochastic parroting in language models, where harmful associations are reproduced and amplified [4]. Text-to-image systems amplify these risks because visual stereotypes can be more immediately visceral and difficult to audit at scale. Simultaneously, the deployment of systems such as Imagen and GLIDE demonstrated that architectural innovations can yield photorealism and compositional reasoning [5, 6], but the same capabilities increase the potential for misuse and representational harm. The present analysis thus situates itself at the intersection of robust alignment and inclusive representation, examining how large-scale infrastructure decisions create path dependencies that constrain subsequent fairness and safety interventions.

2. Background and Related Work

The dominant paradigm for high-quality text-to-image synthesis rests on diffusion models that iteratively denoise random latents conditioned on textual embeddings derived from large language models. While the generative process is well characterized, the vulnerability surface presented by the text encoder has become a focal point for adversarial attacks. Research on extracting training data from diffusion models reveals that even non-adversarial prompts can inadvertently surface memorized visual fragments, raising privacy and copyright concerns [7]. More actively, red-teaming efforts across language and vision-language systems have demonstrated that systematically probing model behaviors with adversarial prompts can uncover failure modes that remain invisible in standard evaluations [8]. The conceptual lineage of adversarial prompts traces back to techniques that identified universal adversarial

triggers capable of manipulating language model outputs across a wide range of inputs [9]. Extending this line of inquiry to the visual domain requires acknowledging that an adversarial prompt need not produce an overtly toxic image to be harmful; it may simply alter subtle stylistic cues that encode ideological messaging or bypass cultural safeguards.

The toxicity of generated content has been studied primarily through the lens of prompt-level benchmarks originally designed for text. The RealToxicityPrompts framework, for instance, provides a systematic methodology to measure the propensity of language models to produce toxic continuations [10]. Analogous efforts for text-to-image outputs remain nascent, in part because toxicity in images is a deeply multidimensional construct that intersects with race, gender, culture, and context. Critical scholarship on multimodal datasets has uncovered deeply embedded misogyny, pornography, and malignant stereotypes that become amplified when models learn to associate certain identity groups with derogatory visual attributes [11]. This scholarship underscores that adversarial prompt vulnerability is not purely a matter of circumventing explicit safety classifiers; rather, it exploits the same distributional biases that shape representation consistency.

Technical approaches to improving alignment in text-to-image systems have included contrastive region guidance, which refines cross-attention mechanisms to enhance spatial correspondence between prompt tokens and image regions whilst mitigating spurious correlations [12]. Personalization techniques that fine-tune models on small sets of user-provided images further demonstrate how embedding spaces can be manipulated through targeted fine-tuning, introducing both opportunities for creative expression and new vectors for adversarial misuse [13]. The environmental footprint of the massive compute required to train and serve these models has also become a subject of concern, with analyses of large language models such as BLOOM quantifying the carbon cost of scaling, and subsequent studies extending these estimates specifically to generative image models like Stable Diffusion [14, 15]. These sustainability considerations intersect with robustness and representation because the energy-intensive nature of retraining or fine-tuning imposes practical limits on how frequently systems can be updated to address newly discovered vulnerabilities or representational gaps.

3. Adversarial Prompt Robustness in Large-Scale Deployments

Adversarial prompt robustness must be understood as a systemic property that emerges from the full stack of a deployed text-to-image system, encompassing the text encoder, the diffusion backbone, the safety classifiers, and the input-output filtering middleware. A narrow focus on hardening the text encoder against lexical perturbations is insufficient because adversaries exploit compositional semantic gaps that lie between the discrete token space and the continuous latent manifold. For instance, an apparently innocuous prompt describing a historical figure can, through subtle phrasing that does not trigger keyword-based filters, guide the model toward generating imagery that encodes harmful stereotypes or glorifies violence. The challenge is compounded by the fact that diffusion models are inherently stochastic, meaning that a slightly perturbed prompt may yield a drastically different output distribution under identical sampling parameters. This stochasticity can be weaponized to perform repeated sampling attacks where an adversary queries the system thousands of times with minor prompt variations until a harmful image slips through the safety net.

The architectural choice of employing a frozen pretrained language encoder, such as the text tower of a contrastive vision-language model, creates a single point of vulnerability. Adversarial perturbations originally developed for text-only models can transfer surprisingly

well to image generation pipelines because the same embedding topology governs the semantic conditioning. Moreover, the scaling of text encoders to billions of parameters, while improving representational fidelity, can also introduce additional adversarial subspaces that are difficult to exhaustively probe during red-teaming. The deployment infrastructure further complicates robustness because production systems often incorporate caching layers, dynamic prompt rewriting, and ensemble-based filtering that can be gamed once an adversary understands their logic. This cat-and-mouse dynamic mirrors well-known patterns in adversarial machine learning for classification, yet the multimodal output space multiplies the degrees of freedom for attackers.

System architects face a structural trade-off between prompt expressiveness and safety. Highly restrictive filtering reduces the attack surface but can also constrain legitimate creative use cases, leading to user dissatisfaction and the rise of unofficial mirror services that lack any safety controls. Conversely, permissive filtering preserves expressiveness but demands continuous, high-frequency monitoring to detect emergent abuse patterns. The sustainability of monitoring regimes is directly tied to operational costs: every adversarial prompt that slips through must be logged, labeled, and fed back into retraining pipelines, incurring both computational and human annotation overhead. The carbon cost of frequent safety-focused fine-tuning cycles, when aggregated across global user bases, can be substantial and must be weighed against the marginal safety gains [14, 15]. This entanglement of adversarial robustness with environmental and economic constraints illustrates why robustness cannot be optimized as an isolated technical metric.

A further dimension of adversarial robustness concerns the integrity of cultural signifiers. Adversaries can deliberately craft prompts that pit two cultural identities against each other, forcing the model into a representational dilemma where any output amplifies a divisive visual narrative. The system's response in such cases reveals whether its alignment training has instilled a robust neutrality or merely suppressed surface-level toxicity. Robustness, therefore, encompasses not only refusal to generate harmful content but also the capacity to generate culturally situated outputs that defuse adversarial intent without sanitizing cultural expression. This form of robustness is inextricably linked to the model's underlying cultural knowledge, a relationship that bridges the gap between adversarial hardening and representation quality.

4. Cultural Representation Consistency and Systemic Bias

Cultural representation consistency in text-to-image systems is a multifaceted challenge that extends beyond simple demographic parity metrics. It encompasses the ability to generate plausible, respectful, and contextually appropriate depictions of clothing, architecture, rituals, landscapes, and social configurations associated with specific cultural groups. When a system systematically fails to produce such imagery, or defaults to Western-centric visual tropes for globally diverse prompts, it participates in what recent scholarship has termed a cultural gap, where the richness of non-mainstream cultures is progressively erased from the generative visual lexicon [16]. This phenomenon is not merely a dataset imbalance issue; it reflects deeper structural decisions about which image-text pairs are prioritized during training, how aesthetic filters bias the notion of visual quality, and whose annotations define the ground truth for concept grounding. The cultural gap can manifest in subtle ways, such as consistently depicting a generic urban street when prompted for a market scene in a Southeast Asian city, or in more overt stereotypical renderings that collapse intra-cultural diversity into a handful of visual clichés.

The origins of cultural inconsistency can be traced to the composition of the training corpora, which are overwhelmingly sourced from English-language web domains and thus encode the visual perspectives of economically dominant regions. Even when multilingual alt-text data are incorporated, the sheer volume of Western imagery dwarfs other cultural sources, leading to a representational hegemony that is resistant to post-hoc debiasing. The process of data filtering further exacerbates the gap. Aesthetic scoring functions, themselves trained on human preference judgments that are culturally skewed, tend to select images that conform to Western photographic conventions in terms of composition, lighting, and color grading. Consequently, the image-text pairs that survive filtering and enter the training pipeline are already biased toward a narrow visual canon, a distortion that propagates through the generative model's latent space.

Architectural choices interact with cultural consistency in non-obvious ways. Models that rely on a single global embedding vector to condition the generative process compress cultural nuance into a high-dimensional point, making it difficult for the model to disentangle cultural attributes from correlated spurious features such as geographical region or socioeconomic status. Contrastive learning objectives that align image and text representations can inadvertently penalize cultural specificity if such specificity is not well represented in the training distribution, since the model learns an averaged semantic space that smooths over minority variations. The use of classifier-free guidance, while beneficial for sample quality, has been observed to amplify dominant modes of the distribution, potentially exacerbating the underrepresentation of cultural minorities by suppressing low-probability visual tokens that are precisely those that carry cultural distinctiveness.

Addressing cultural representation consistency demands more than incremental data augmentation. It requires culturally grounded annotation efforts that involve community knowledge holders, an approach that challenges the dominant paradigm of crowdsourced labeling by anonymous workers. The development of culturally stratified evaluation benchmarks that go beyond coarse geographic bins is an essential infrastructural investment. Furthermore, the deployment of fine-grained controllable generation interfaces could allow users to specify cultural parameters, but such mechanisms must be designed to prevent their use as tools for stereotyping or cultural appropriation. The tension between user control and representational safety mirrors the expressiveness-robustness trade-off discussed earlier, highlighting that adversarial prompt robustness and cultural consistency are two facets of a single alignment challenge. A system that is robust against adversarial prompts that attempt to provoke cultural offense must possess a sufficiently nuanced cultural representation to recognize and defuse such prompts without falling back on generic refusals that themselves convey cultural indifference.

5. Infrastructure, Architecture, and Governance Trade-offs

The simultaneous pursuit of adversarial prompt robustness and cultural representation consistency introduces a set of deep structural trade-offs that play out across the layers of system architecture. At the model architecture level, the choice between a monolithic text encoder shared across all tasks and a modular design with specialized encoders for safety and cultural context is a central design decision. A monolithic encoder is computationally efficient and simplifies training, but it tightly couples the representations used for creativity, safety, and cultural grounding, making it difficult to update one without destabilizing the others. A modular approach, where a dedicated prompt safety encoder operates in parallel with the primary conditioning pathway, allows for targeted adversarial hardening, but it increases

inference latency and energy consumption, and introduces interfaces that can become new adversarial surfaces if the coordination logic is not carefully designed.

The training infrastructure for large-scale text-to-image systems is another locus of trade-offs. End-to-end training from scratch on meticulously curated culturally balanced datasets is prohibitively expensive and would likely result in a lag in terms of overall image quality compared to continual training on web-scale data. The prevailing practice of fine-tuning pretrained models on auxiliary safety and representation objectives creates a path-dependent trajectory where the initial pretraining distribution's biases become locked into the latent manifold, and subsequent fine-tuning can only partially overwrite them. This phenomenon underscores the need for pretraining data governance frameworks that embed cultural audit processes at the data acquisition stage rather than treating representation as a downstream correction task.

Safety-oriented middleware, including prompt classifiers, output detectors, and human-in-the-loop moderation queues, constitutes yet another critical layer. The operational logic of these components often operates on a binary harmful/benign axis that is poorly suited to capturing the gradations of cultural insufficiency. A prompt requesting an image of a traditional wedding ceremony is not harmful, but if the model reliably produces a Western-style wedding regardless of cultural specification, the output is representationally inadequate, a failure mode that escapes toxicity classifiers. Consequently, infrastructure monitoring must evolve toward multidimensional quality signals that jointly assess safety, cultural fidelity, and diversity. Such monitoring systems generate continuous streams of telemetry that must be interpreted through governance protocols defining thresholds for forced retraining, model rollback, or content suppression.

Governance arrangements for adversarial prompt robustness and cultural consistency are themselves subject to trade-offs between centralized control and distributed oversight. Centralized safety teams can enforce consistent standards but often lack the cultural competencies to adjudicate edge cases involving non-Western contexts. Community-based feedback loops, wherein cultural organizations contribute to model audits and red-teaming exercises, improve cultural accuracy but introduce coordination costs and potential conflicts over representational authority. A polycentric governance model, in which multiple overlapping institutions set standards, share audit data, and negotiate remediation actions, appears promising for managing the complexity of global deployments, though it demands robust infrastructure for transparency reporting and inter-organizational trust. The procedural legitimacy of such governance depends on whether affected communities have genuine agency in shaping the representation criteria, a challenge that reverberates down to the technical level of specifying loss functions and evaluation metrics.

6. Deployment, Policy, and Sociotechnical Implications

The deployment of text-to-image systems at planetary scale transforms questions of adversarial prompt robustness and cultural representation consistency from laboratory research problems into sociotechnical policy challenges with immediate societal consequences. When a single model produces millions of images daily, even a miniscule rate of adversarial prompt success can translate into a torrent of harmful visual content, while a consistent pattern of cultural erasure compounds over time into a form of algorithmic epistemicide that shapes collective visual memory. Policy responses must therefore address the temporal dimension of harm accumulation, moving beyond incident-level remediation toward systemic accountability frameworks. Model cards and similar transparency

instruments provide a mechanism for documenting intended use, training data composition, and evaluation results [18], yet they currently lack standardized metrics for adversarial robustness over time and for cultural representation fidelity across disaggregated identity axes.

Regulatory initiatives that emphasize mandatory risk assessments and third-party audits for foundation models create an opportunity to formalize adversarial prompt robustness and cultural consistency as auditable properties. However, the practical implementation of such audits requires the development of open, maintainable benchmark suites that capture the adversarial dynamics of prompt evolution and the cultural diversity of global users. The compilation of these benchmarks must itself be guarded against cultural appropriation, ensuring that the act of curating a culturally diverse test set does not reduce living traditions to static token sequences. The release strategies of generative models, whether gradual staged deployment or open release, profoundly influence the adversarial discovery rate and the diffusion of representational harms [21]. Open releases have democratized access and enabled grassroots innovation, but they have also made it dramatically harder to retrospectively impose safety guardrails once community fine-tuned variants proliferate.

The sociotechnical lens further reveals that adversarial prompt attacks and cultural representation failures are often entwined with economic incentives. Adversarial prompt marketplaces that sell jailbreaking techniques monetize the robustness gaps, creating a commercial ecosystem that thrives on the lag between vulnerability discovery and patching. Similarly, the cultural gap can be exploited to generate cliché-ridden stock imagery that reinforces stereotypes, generating revenue streams that disincentivize the costly curation of diverse training data. Breaking these incentive structures may require policy interventions such as algorithmic content provenance standards and mandatory cultural impact disclosures that make the representational costs visible to consumers and investors.

The evaluation of object recognition systems across geographic and income strata has demonstrated that representational harms are not uniformly distributed, with performance disparities disproportionately affecting underrepresented regions [22]. Text-to-image generation inherits and amplifies these disparities because the visual knowledge required to depict objects and scenes from non-Western contexts is systematically attenuated. Architectures that hierarchically condition generation on language latents, such as the two-stage pipeline employed in some scalable production systems, can decouple semantic planning from pixel synthesis, potentially offering more granular control over the rendering of culturally specific details [23]. The diffusion process itself can be steered through classifier-free guidance, which modulates the influence of the unconditional model to control fidelity, yet this mechanism, if tuned solely on aggregate aesthetic preferences, risks further suppressing rare visual modes that are essential for cultural diversity [24]. Addressing these structural tendencies requires not only algorithmic adjustments but a fundamental rethinking of the relationship between technical abstraction and social context, as highlighted by scholarship on fairness in sociotechnical systems [25].

7. Conclusion

Adversarial prompt robustness and cultural representation consistency have emerged as dual imperatives for the next generation of large-scale text-to-image systems. This paper has argued that these properties cannot be meaningfully addressed as independent technical problems; they are co-constituted by the architectural decisions, data supply chains, moderation pipelines, and governance structures that define the sociotechnical infrastructure of generative AI. The trade-offs between expressiveness and safety, between global

homogenization and cultural specificity, and between centralized oversight and community participation demand an integrated analytical frame. Progress will require the cultivation of evaluation ecosystems that treat robustness and representation as dynamic, context-dependent qualities rather than static checkboxes. The energy and economic costs of continuous monitoring and retraining must be acknowledged within sustainability planning, ensuring that safety and fairness do not become an environmental burden disproportionally borne by marginalized communities. Future research trajectories should investigate modular architectural designs that allow independent updates to safety and cultural grounding components, culturally informed red-teaming methodologies that go beyond keyword-based attack generation, and governance frameworks that vest real decision-making power in the communities most affected by representational harms. Ultimately, building trustworthy text-to-image infrastructure is not a destination but a continuous process of negotiation among technological capability, cultural plurality, and democratic accountability.

References

1. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (pp. 8748–8763). PMLR.

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

3. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

4. 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). ACM.

5. 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, 35, 36479–36494.

6. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... & Chen, M. (2021). GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741.

7. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., ... & Wallace, E. (2023). Extracting training data from diffusion models. In 32nd USENIX Security Symposium (pp. 5253–5270). USENIX.

8. Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., ... & Irving, G. (2022). Red teaming language models with language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 3419–3448). ACL.

9. Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal adversarial triggers for attacking and analyzing NLP. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (pp. 2153–2162). ACL.

10. Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 3356–3369). ACL.

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

12. Cho, J., Zala, A., & Bansal, M. (2023). Contrastive region guidance: Improving text-to-image alignment without additional training. arXiv preprint arXiv:2303.02325.

13. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using a single image. arXiv preprint arXiv:2203.12693.

14. Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2023). Estimating the carbon footprint of BLOOM, a 176B parameter language model. Journal of Machine Learning Research, 24(253), 1–15.

15. Luccioni, A. S., Lacoste, A., & Schmidt, V. (2023). Estimating the carbon footprint of generative AI: A case study on Stable Diffusion. arXiv preprint arXiv:2311.16863.

16. C. Shi, S. Li, S. Guo, S. Xie, W. Wu, J. Dou, C. Wu, C. Xiao, C. Wang, Z. Cheng, et al. (2025)Where culture fades: revealing the cultural gap in text-to-image generation.arXiv preprint arXiv:2511.17282.

17. 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.

18. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). ACM.

19. 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. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44). ACM.

20. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

21. Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., ... & Wang, J. (2019). Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203.

22. 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). IEEE.

23. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125.

24. Ho, J., & Salimans, T. (2021). Classifier-free diffusion guidance. In NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications.

25. 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). ACM.