Federated Auditing of Cultural Fairness in Privacy-Preserving Text-to-Image Generation Models

Ross Franklin
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
rossfranklin@ucf.edu

Shanae D. Rese
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
srose@missouri.edu

Hudson A. Nieminen
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
hudsonmail@buffalo.edu

Abstract

The rapid proliferation of text-to-image generative models has surfaced deep structural concerns regarding cultural fairness, representation, and inclusion. These models, trained on vast and often uncurated web-scale datasets, exhibit significant cultural biases that risk homogenizing visual expression and marginalizing underrepresented communities. Simultaneously, regulatory frameworks and community expectations increasingly demand rigorous auditing of such systems without compromising the data sovereignty of the cultural groups involved. This paper proposes a federated auditing architecture that enables the systematic evaluation of cultural fairness in text-to-image models while preserving the privacy and local control of culturally sensitive data. We present a system-level analysis of the trade-offs inherent in distributing fairness auditing across heterogeneous cultural nodes, discussing architectural design, secure aggregation protocols, governance structures, and policy alignment. Through a detailed conceptual framework, we examine how federated fairness auditing can reconcile the competing demands of statistical rigor, cultural authenticity, and privacy preservation. The discussion extends to deployment infrastructure, robustness against adversarial manipulation, and the integration of community-defined fairness metrics. By centering the infrastructural and governance dimensions, this work provides a comprehensive pathway toward accountable and culturally aware generative artificial intelligence systems, highlighting the necessity of interdisciplinary collaboration between systems engineers, cultural stakeholders, and policy makers.

Keywords

Federated Learning, Cultural Fairness, Text-to-Image Models, Privacy-Preserving Auditing, Algorithmic Fairness, Decentralized Governance.

1. Introduction

The emergence of highly capable text-to-image generative models, exemplified by latent diffusion architectures and large-scale transformer-based systems, has transformed creative production and digital media synthesis [1]. These models translate natural language prompts into visually coherent and semantically rich images by learning from billions of image-text

pairs, enabling applications ranging from content creation to assistive design. Alongside their technical prowess, however, significant societal questions have emerged regarding the cultural dimensions of their outputs. Fairness and bias in machine learning have been extensively studied in classification and decision-making contexts [2], but generative models introduce distinct challenges, as they actively produce rather than merely reflect cultural representations. The cultural values embedded in the model's latent space often default to dominant Western, Anglophone, and urban-centric aesthetics, inadvertently erasing or distorting the visual idioms of minority and indigenous cultures.

In parallel, the imperative to protect sensitive user data and respect data sovereignty has driven the widespread adoption of privacy-preserving techniques such as federated learning, which enables collaborative model improvement without centralizing raw data [3]. While federated learning has been predominantly applied to model training, its principles are equally relevant to the auditing of model behavior, particularly when auditing data is culturally sensitive or legally restricted. The ability to evaluate fairness without extracting cultural artifacts from their local contexts aligns with broader movements toward data sovereignty and community-driven AI governance. Recent investigations have specifically documented pervasive cultural biases in generative image synthesis, revealing that prompts tied to non-Western cultural contexts frequently yield stereotyped, inaccurate, or simplistic visual outputs [4]. These findings underscore the urgency of developing auditing mechanisms that are both culturally informed and technically robust. Furthermore, indigenous data sovereignty frameworks argue that communities must retain authority over how their cultural data is used, analyzed, and represented [5]. This principle stands in tension with centralized auditing paradigms that require aggregating culturally specific datasets into a single repository, motivating the exploration of federated alternatives.

The present work proposes and analyzes a federated architecture for auditing cultural fairness in privacy-preserving text-to-image generation models. We conceptualize cultural fairness not as a monolithic metric but as a multifaceted property that emerges from the alignment between model outputs and the self-defined representational norms of distinct cultural communities. Our system-level analysis addresses the structural trade-offs between decentralization, privacy guarantees, fairness fidelity, and operational scalability. By integrating perspectives from distributed systems, cultural theory, and AI governance, we articulate a comprehensive vision for accountable generative AI that respects cultural plurality while maintaining technical rigor.

2. Background and Related Work

Cultural bias in computer vision and language models has been recognized as a critical challenge arising from the skewed geographic and demographic distribution of training data. Early work highlighted geodiversity gaps in open image datasets, demonstrating that Western-centric imagery dominates standard benchmarks, leading to systematic underrepresentation of vast regions and populations [6]. This data imbalance directly propagates into generative models, which not only replicate but often amplify existing cultural stereotypes. The challenge goes beyond geographical representation, extending into the nuanced semiotics of cultural symbols, attire, rituals, and artistic styles that are encoded and often flattened by deep learning systems. Studies on cultural bias in natural language processing have similarly revealed that language models routinely fail to capture the complex social and cultural factors that shape meaning, posing risks of misrepresentation when textual prompts interface with

visual generation [7]. These cultural blind spots are not merely technical artifacts but reflect deep structural biases in data collection, annotation, and model design paradigms.

The technical foundation for privacy-preserving distributed computation has been established through differential privacy and secure aggregation methods. Differential privacy provides a formal framework for bounding the privacy loss associated with any individual data record, typically through the calibrated injection of noise into query responses or model updates [8]. In the context of auditing, differential privacy could protect the sensitive cultural indicators embedded in evaluation datasets while still allowing the extraction of statistically meaningful fairness signals. Concurrently, the specific cultural dimension of text-to-image generation has been empirically quantified in recent work that systematically revealed the cultural gap across major generative platforms, showing that prompts referencing non-Western cultures frequently result in stereotypical, inaccurate, or aesthetically impoverished images while Western prompts produce rich, diverse outputs [9]. This finding provides a crucial empirical baseline that motivates the need for continuous and culturally granular auditing.

Federated learning has evolved from its initial communication-efficient learning formulation to encompass a broad set of protocols for decentralized model training and evaluation. Secure aggregation protocols ensure that a central server can compute the sum or average of client model updates without inspecting individual contributions, thus preserving privacy in honest-but-curious settings [10]. Extensions to Byzantine-robust aggregation provide resilience against clients that deviate arbitrarily from the prescribed protocol, whether due to adversarial intent or systemic faults [11]. Such robustness is especially critical in cultural fairness auditing, where a malicious actor might attempt to skew fairness assessments to conceal biased behavior. On the auditing front, algorithmic fairness auditing has developed into a mature discipline, with frameworks emphasizing the importance of end-to-end organizational accountability, documentation, and rigorous statistical testing [12]. These frameworks, however, largely assume centralized data access, which is incompatible with the requirements of cultural data sovereignty. Federated fairness evaluation has recently been proposed to bridge this gap, outlining how fairness metrics can be estimated without direct access to raw data through secure computation and appropriately designed sampling [13]. Closely related, privacy-preserving auditing architectures leverage differential privacy and secure multi-party computation to enable black-box model assessments under formal privacy guarantees [14]. The intersection of these lines of work with cultural heritage applications further motivates architectures that can handle culturally sensitive data with the requisite ethical and legal care [15].

3. System Architecture for Federated Cultural Fairness Auditing

The proposed federated auditing architecture is built upon three core components: a set of geographically and culturally distributed client nodes, a coordination server responsible for orchestrating the auditing protocol, and a secure communication substrate that enforces privacy and integrity guarantees. Each client node represents a distinct cultural community, holding a curated dataset of culturally representative prompts and corresponding ground-truth annotations regarding expected visual attributes, stylistic conventions, or symbolic correctness. These datasets are never transmitted to the server; instead, each node locally queries the text-to-image model under audit using its own prompt set, records the generated images, and computes a vector of locally defined fairness metrics. These metrics may include representation accuracy, stereotype prevalence, cultural token fidelity, and diversity indices,

all of which are aligned with community-defined values rather than externally imposed universal standards.

The coordination server aggregates the local metric vectors using a secure aggregation protocol, ensuring that the server learns only the aggregate statistics without access to any individual community's metrics or raw data. This design leverages advances in secure aggregation [10] and differential privacy [8], adding calibrated noise at the client level to obscure individual contributions while still enabling the computation of population-level fairness estimates with quantifiable uncertainty. Privacy budget accounting occurs per auditing round, allowing communities to enforce strict limits on information leakage over time. The architecture must accommodate significant heterogeneity in metric design, dataset size, and computational resources across clients. A naive averaging of locally computed fairness scores could obscure critical disparities, especially when small or marginalized communities produce high-variance estimates due to limited data. Therefore, the aggregation protocol incorporates variance-aware weighting and can leverage hierarchical federated structures where regional clusters of culturally similar communities jointly compute intermediate aggregates before global coordination. This hierarchical approach both improves statistical efficiency and respects cultural affinities by allowing intermediate groupings that reflect shared cultural traits or linguistic ties.

Critical system trade-offs emerge between the granularity of the cultural fairness assessment and the strength of the privacy guarantees. Finer-grained auditing that requests more detailed statistics or multiple condition-specific metrics from each client inevitably consumes more privacy budget and increases communication costs. The architecture must therefore support an adaptive auditing cadence, where routine lightweight audits are complemented by periodic deep audits that expend higher privacy budgets under explicit community consent. The system design further requires that the auditing infrastructure be extensible to different text-to-image model providers, necessitating standardized interfaces for model querying and metric reporting. The federated auditing framework can be deployed as a middleware layer between model providers and cultural communities, enabling plug-and-play integration with existing generative model APIs while maintaining architectural coherence.

An equally important architectural dimension is the handling of distributional shifts in model outputs over time. Text-to-image models are frequently updated through fine-tuning, reinforcement learning from human feedback, or dataset expansions, all of which can alter cultural representations. Federated auditing must therefore operate incrementally, efficiently detecting changes in fairness metrics without requiring full re-audits from all communities. Incremental federated auditing protocols based on streaming differential privacy and change-point detection can provide continuous oversight with minimal bandwidth, alerting stakeholders to regressions in cultural fairness shortly after model updates occur. The server's role extends beyond aggregation to include dashboards, anomaly detection, and the issuance of fairness certificates that model providers can use to demonstrate compliance with cultural representation standards.

4. Governance and Cultural Representation

The technical architecture of federated auditing cannot succeed without a robust governance model that determines who defines cultural fairness, how community representation is established, and what mechanisms exist for dispute resolution. Cultural fairness is inherently plural and contested; what constitutes an authentic or respectful representation of a cultural motif is not a fixed property but is negotiated within communities and across generations.

Therefore, the governance framework must empower cultural communities to define their own auditing metrics and evaluation protocols rather than imposing a universal rubric. This requires a multi-stakeholder governance body that includes cultural leaders, domain anthropologists, legal experts, and technologists, collectively responsible for validating the authenticity and ethical grounding of client-side auditing procedures. Indigenous data sovereignty principles, which assert the right of indigenous peoples to govern the collection, ownership, and application of data about their communities, offer a compelling doctrinal foundation for such governance [5].

Representation in the auditing ecosystem must be both inclusive and procedurally legitimate. Small, stateless, or diaspora communities must have pathways to participate as auditing clients with equal voting weight in governance decisions, even if their technical infrastructure is limited. This may require a subsidized client deployment model, where international organizations or public-interest technology funds provide the necessary computational and networking resources. The governance framework must also address the dynamic nature of cultural identity, allowing for the emergence of new cultural collectives and the evolution of existing ones without bureaucratic obstruction. From a systems perspective, this translates into a dynamic client registration protocol with lightweight identity verification that respects community self-definition while preventing Sybil attacks that could flood the auditing network with fabricated nodes. Techniques from decentralized identity and verifiable credentials can be employed to establish trust without requiring centralized state-based authentication. Regular audits of the governance process itself become necessary to detect capture by dominant cultural actors or economic interests that might seek to dilute the representation of marginalized voices.

Regulatory alignment further shapes the governance architecture. The European Union's Artificial Intelligence Act establishes requirements for high-risk AI systems to undergo conformity assessments that include fairness and non-discrimination criteria [16]. While the Act does not explicitly mandate cultural fairness auditing, its provisions for transparency, human oversight, and risk management create a regulatory pull toward systematic evaluation of generative model outputs. Similarly, the General Data Protection Regulation's principles of data minimization and purpose limitation reinforce the need for privacy-preserving approaches that do not require centralization of personal or culturally sensitive data [17]. Federated auditing aligns with these regulatory trends by enabling evidence-based compliance without violating data protection mandates. Governance mechanisms should produce audit trails that are admissible in regulatory proceedings, necessitating cryptographic commitments and tamper-evident logs that record the sequence of auditing rounds, client participation, and aggregate results.

5. Infrastructure and Deployment Considerations

Deploying a federated cultural fairness auditing system at scale introduces substantial infrastructure challenges that span networking, heterogeneity management, and reliability. The client nodes may range from well-resourced data centers in urban hubs to low-power edge devices in remote cultural communities. This heterogeneity demands a communication protocol that is resilient to intermittent connectivity, high latency, and bandwidth constraints. Asynchronous federated aggregation schemes, where the server proceeds with available client updates without waiting for stragglers, become necessary to avoid indefinite delays caused by clients in underserved regions. However, asynchronous aggregation introduces staleness in

fairness estimates, requiring compensation mechanisms such as importance weighting based on update recency and network quality metrics.

The server infrastructure must be designed for high availability and must withstand targeted denial-of-service attacks that could disrupt fairness auditing during critical regulatory windows. Replicated state machine techniques and geographically distributed server replicas can ensure that the auditing process remains operational even under adversarial conditions. The economic model is equally infrastructural: who pays for the compute and network resources consumed by auditing clients? A sustainable deployment could adopt a model where model providers fund the auditing infrastructure as part of their regulatory compliance obligations, with neutral audit foundations operating the server layer to prevent conflicts of interest. The client nodes, particularly those serving marginalized communities, should receive financial support without strings attached, preserving their independence in defining fairness metrics. The technical substrate can include automated micropayment channels based on smart contracts, compensating clients for each completed auditing round, thereby creating economic incentives for sustained participation.

Robustness against Byzantine behavior forms a critical layer of the infrastructure. A malicious client might submit deliberately skewed fairness reports designed to inflame inter-community tensions or to mask actual biases in a colluding model provider's outputs. Byzantine-resilient aggregation methods, such as coordinate-wise median and trimmed mean approaches [11], can mitigate the impact of such outliers. However, applying these methods directly to fairness metrics is nontrivial, as cultural fairness scores are often bounded and exhibit non-Gaussian distributions. More sophisticated robust aggregation that leverages pairwise similarity between clients' metric vectors and historical correlation structures can provide better detection and suppression of anomalous contributions. The system can also incorporate reputation scores that evolve based on the consistency of a client's reports with corroborating evidence from indirect model probing, although care must be taken to avoid unjustly devaluing legitimate but minority cultural viewpoints that naturally diverge from majority norms. The tension between outlier robustness and epistemic pluralism constitutes a fundamental design challenge that demands ongoing interdisciplinary research and adaptive system parameters.

6. Policy Implications and Future Directions

The adoption of federated cultural fairness auditing holds profound implications for AI policy and international cultural governance. It operationalizes the principle of cultural self-determination in algorithmic systems, shifting the locus of evaluation from centralized corporate or state authorities to distributed community networks. This shift aligns with emerging global norms on digital cultural rights and the protection of intangible cultural heritage in the digital age. Policymakers can leverage federated auditing outputs to establish enforceable fairness thresholds in public procurement of generative AI services, conditioning government use of such models on verified cultural fairness scores. International treaties on cultural diversity, such as the UNESCO Convention on the Protection and Promotion of the Diversity of Cultural Expressions, could be modernized to incorporate provisions for algorithmic accountability, with federated auditing serving as an implementation mechanism. Auditing infrastructure could ultimately be recognized as a form of digital public infrastructure, akin to digital identity or payment systems, warranting investment from multilateral development banks and philanthropic foundations.

Future technical work must develop formal methods for composing local cultural fairness metrics into globally meaningful indices that do not erase local nuance. This requires advances in nonparametric statistics for federated settings and new privacy accounting techniques that handle adaptive queries over time. The integration of federated auditing with model alignment pipelines presents another frontier: audit signals could directly inform model fine-tuning through differentially private feedback loops, enabling continuous cultural calibration. From a legal standpoint, the evidentiary weight of federated audit reports must be established in courts and regulatory proceedings, demanding standardization of protocols and independent certification of auditing software and hardware stacks. Interdisciplinary collaborations that bring together systems engineers, ethnographers, legal scholars, and community organizers are essential to co-design the next generation of culturally aware and technically robust auditing frameworks. The path forward lies in recognizing that cultural fairness cannot be engineered solely in the laboratory; it must be woven into the fabric of socio-technical systems through distributed, participatory, and privacy-respecting architectures.

7. Conclusion

This paper has presented a comprehensive system-level analysis of federated auditing as a mechanism for evaluating cultural fairness in privacy-preserving text-to-image generation models. We have argued that centralized auditing paradigms are fundamentally incompatible with the principles of cultural data sovereignty and the heterogeneous nature of cultural fairness itself. The federated architecture we propose leverages secure aggregation, differential privacy, and Byzantine-robust techniques to enable communities to assess generative model outputs according to their own definitions of representational fidelity, without surrendering sensitive cultural data. We analyzed the structural trade-offs between audit granularity, privacy intensity, and system scalability, highlighting how adaptive auditing cadences and hierarchical aggregation can reconcile these tensions. Governance, infrastructure, and policy dimensions were examined as integral components, not mere afterthoughts, of a viable auditing ecosystem. The framework's alignment with emerging AI regulations and its potential to serve as digital public infrastructure underscore its relevance to both technology development and cultural policy. As generative AI systems become increasingly pervasive in shaping visual culture, federated cultural fairness auditing offers a pathway toward a more pluralistic digital future where all communities have the agency to define, monitor, and uphold their representational integrity.

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, 10684–10695.

2. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

3. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273–1282.

4. Luccioni, A. S., & Rolnick, D. (2023). Stable bias: Analyzing societal representations in diffusion models. arXiv preprint arXiv:2303.11408.

5. Kukutai, T., & Taylor, J. (Eds.). (2016). Indigenous data sovereignty: Toward an agenda. ANU Press.

6. Shankar, S., Sivakumar, V., Mallya, A., Yu, F. X., & Ravi, S. (2017). No classification without representation: Assessing geodiversity issues in open data sets for computer vision. Workshop on Fairness, Accountability, and Transparency in Machine Learning, NeurIPS.

7. Prabhakaran, V., Qadri, R., & Hassine, T. (2022). Cultural and geographical biases in language models. arXiv preprint arXiv:2212.13785.

8. Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Proceedings of the Third Theory of Cryptography Conference, 265–284.

9. 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.

10. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175–1191.

11. Blanchard, P., Guerraoui, R., Stainer, J., et al. (2017). Machine learning with adversaries: Byzantine-tolerant gradient descent. Advances in Neural Information Processing Systems, 30.

12. 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 Conference on Fairness, Accountability, and Transparency, 33–44.

13. Shi, Y., Yu, H., & Leung, C. (2023). A survey on fairness in federated learning. ACM Computing Surveys, 55(2), 1–38.

14. Panchamukhi, P., Mahajan, G., & Ananthanarayanan, G. (2022). Towards federated fairness auditing. arXiv preprint arXiv:2208.05115.

15. Margetis, G., Ntoa, S., Antona, M., & Stephanidis, C. (2021). AI for cultural heritage: Challenges and perspectives. Pattern Recognition Letters, 145, 128–136.

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

17. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. International Data Privacy Law, 7(2), 76–99.

18. Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. Proceedings of the 35th International Conference on Machine Learning, 5640–5649.

19. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308–318.

20. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210.