arXiv:2505.23814v1 [cs.CR] 27 May 2025

Watermarking Without Standards Is Not AI Governance

Alexander Nemecek, Yuzhou Jiang, Erman Ayday
Case Western Reserve University
`{ajn98, yxj466, exa208}@case.edu`

Abstract

Watermarking has emerged as a leading technical proposal for attributing generative AI content and is increasingly cited in global governance frameworks. This paper argues that current implementations risk serving as symbolic compliance rather than delivering effective oversight. We identify a growing gap between regulatory expectations and the technical limitations of existing watermarking schemes. Through analysis of policy proposals and industry practices, we show how incentive structures disincentivize robust, auditable deployments. To realign watermarking with governance goals, we propose a three-layer framework encompassing technical standards, audit infrastructure, and enforcement mechanisms. Without enforceable requirements and independent verification, watermarking will remain inadequate for accountability and ultimately undermine broader efforts in AI safety and regulation.

1 Introduction

> *"A law without teeth is just a suggestion."*

This adage captures a growing concern in AI governance, where policies are advancing more quickly than the technical tools available to enforce them. A prominent example of this mismatch is *watermarking*, a family of techniques designed to embed identifiable signatures into AI-generated content. Watermarking is widely referenced in AI governance frameworks, yet current implementations often lack the robustness, transparency, and verifiability required to fulfill this role effectively.

Recent advances in generative AI have significantly increased the scale and realism of synthetic content, including text, images, and audio [26, 13]. As such content is deployed in sensitive areas like education, healthcare, and finance, policymakers have identified attribution and provenance as urgent challenges. Misattributed content contributes to the spread of misinformation and to technical failures such as feedback loops in model training, including contamination and eventual model collapse [25].

In response, watermarking has gained traction in both technical research [17, 12, 10] and policymaking. The United States Executive Order 14110, for example, mandates "state-of-the-art" provenance tools and cites watermarking explicitly [1]. The European Union's AI Act requires machine-readable content markings [2], with similar provisions being proposed in jurisdictions worldwide [34]. Although watermarking is just one of several provenance strategies, it occupies an important role in current governance discourse.

This reliance rests on a flawed foundation. While watermarking techniques vary across modalities, most remain brittle, difficult to audit, and proprietary. Policymakers often assume these methods can be standardized and verified. In practice, industry deployments obscure technical details while asserting compliance, turning watermarking into a box-checking exercise rather than a meaningful

Preprint. Under review.

tool. Lacking common standards, evaluation infrastructure, or defined threat models, current implementations are unlikely to fulfill their intended governance role.

To address this growing disconnect, **we advance two central positions:**

- **Watermarking schemes must be designed with verifiability and auditability as primary technical requirements, rather than implemented as proprietary black boxes.**
- **Policymakers must establish technical standards and independent testing to ensure that watermarking fulfills governance goals in practice.**

To support these positions, we make three contributions. First, we analyze emerging watermarking mandates and demonstrate they assume levels of technical feasibility that current systems do not meet. Second, we examine the incentive structures driving industry watermarking deployments, showing how market dynamics often disincentivize robust, auditable implementations. Third, we propose a three-layer framework, spanning technical design, audit infrastructure, and enforcement which realigns the design of watermarking systems with their intended governance functions.

While prior work has analyzed watermarking's technical constraints, our contribution reframes these issues through a governance lens and provides a structured, actionable framework for aligning technical design with regulatory intent. Without intervention at the intersection of policy and design, watermarking is unlikely to deliver meaningful accountability. Instead, it risks becoming a symbolic substitute for the more demanding components of effective AI regulation.

2 Governance Aspirations vs. Technical Realities

As watermarking becomes a feature of AI governance proposals, many mandates rely on technical assumptions that current systems cannot fulfill. We identify three such assumptions and contrast each policy aspiration with technical realities, highlighting key gaps that must be addressed for watermarking to serve as an effective governance tool.

2.1 Assumption I: Watermarking Is Technically Robust Against Modification

Many governance proposals assume watermarking is robust enough to survive benign transformations and adversarial tampering. Whether applied to text, images, or audio, watermarking is expected to remain detectable throughout the content lifecycle.

> *"The disclosure is permanent or extraordinarily difficult to remove, to the extent it is technically feasible."* — California Senate Bill No. 942 [9]

> *"Such techniques and methods should be sufficiently reliable, interoperable, effective and robust as far as this is technically feasible, taking into account available techniques or a combination of such techniques, such as watermarks,... as may be appropriate."* — European Union AI Act [2]

These documents reflect the shared policy expectation that watermarking should resist removal or degradation in any environment. None define quantitative thresholds, making compliance and enforcement difficult. Additionally, no shared common metrics exist for evaluating watermark robustness under adversarial conditions.

In practice, watermarking schemes often fall short of this robustness ideal. Google's SynthID [15], for example, embeds imperceptible watermarks in AI-generated text. While resilient to certain distortions, its detection can be evaded by simple edits such as character perturbations or short-form text paraphrasing [11, 19]. Across all modalities, there is currently no standardized evaluation protocol for watermark robustness, and few systems offer guarantees under adversarial or worst-case conditions.

The absence of shared benchmarks or test suites for robustness creates two problems: first, it makes meaningful regulatory compliance unverifiable; second, it enables firms to implement weak watermarking schemes while claiming alignment with policy language. To concretely illustrate this gap in industry, we conducted a minimal text robustness evaluation using the open-source MarkLLM framework [21], configured with Google's SynthID-Text [11] watermarking on `Gemma-2-9b-it` [14]. We selected this model because it reflects Google's public generative model family, with `2` being a relatively recent version, `9b` offering a balance between capability and accessibility, and `it` indicating instruction tuning, representing realistic deployment settings.

We generated outputs for five prompts spanning factual, narrative, and informal content types, and evaluated three variants of each: the original, a paraphrased version, and a summarized version. Transformations were applied using GPT-4o [16], reflecting a realistic scenario where users reprocess LLM outputs. Detection scores and classifications were obtained from SynthID with a threshold of 0.52 from MarkLLM. As shown in Table 1, even benign edits frequently degraded detection scores below this threshold. Full prompt examples and transformed outputs are provided in Appendix A.

Table 1: SynthID-Text detection results for five prompts across three transformations: original model output, paraphrasing, and summarization. A threshold of 0.52 is used for classification, with values ≥ 0.52 labeled as **True** (watermark detected).

Prompt	**Transformation**	**Score**	**Detected**
Prompt$_1$	Original	**0.5201**	**True**
	Paraphrase	0.4894	False
	Summarize	0.5145	False
Prompt$_2$	Original	**0.5267**	**True**
	Paraphrase	0.4934	False
	Summarize	0.4969	False
Prompt$_3$	Original	**0.5214**	**True**
	Paraphrase	0.5176	False
	Summarize	0.5129	False
Prompt$_4$	Original	0.5054	False
	Paraphrase	0.5008	False
	Summarize	0.5111	False
Prompt$_5$	Original	**0.5261**	**True**
	Paraphrase	0.5092	False
	Summarize	0.4869	False

2.2 Assumption II: Watermarking Enables Independent and Reliable Detection

A second assumption in many governance proposals is that watermarking will enable independent third parties (e.g., regulators, researchers, platform operators) to reliably detect AI-generated content. This assumption seeks to enforce provenance or hold actors accountable for synthetic content.

> *"[Develop] Methods for the verification of statements of digital content provenance to ensure authenticity such as watermarking or classifiers, which are trained models that distinguish artificial intelligence-generated content."* — United States Congress S. 3312 [28]

> *"Deep synthesis service providers shall fulfill entity responsibility for information security, establish and complete management structures for algorithmic mechanism and logic examination and verification... with secure and controllable technical safeguard mechanisms suited to the development of new technologies and new applications."* — China's Deep Synthesis Provisions [8]

These statements imply that watermarking should support reliable, externally verifiable detection. Most watermarking systems today do not support this expectation due either to limited transparency or configuration dependence.

For instance, OpenAI previously released a classifier for detecting AI-generated text, but withdrew it due to unreliability [20]. Google's SynthID offers a more complete approach, with watermarking and detection tools for images and text. SynthID-Text [11] supports watermarking and detection in public tools, but detection still requires access to the configuration used during generation. Additionally, user's need to apply to receive access to the main configurations used in deployment. Unless watermarking keys or models are shared or standardized, third-party detection remains limited to contexts explicitly designed for interoperability.

While the technical foundations for third-party detection are emerging, the practical reality remains constrained. Most watermarking approaches do not currently support universal, auditable detection, and governance frameworks often overlook the infrastructure and standardization needed to make detection viable at scale.

2.3 Assumption III: Industry Will Voluntarily Align with Governance Goals

A final assumption in governance frameworks, particularly in the United States and European Union, is that generative AI providers will voluntarily adopt watermarking practices aligned with policy goals.

> *"... establish a strong international framework for managing the risks and harnessing the benefits of AI, including by encouraging international allies and partners to support voluntary commitments similar to those that United States companies have made in pursuit of these objectives... to develop common regulatory and other accountability principles for foreign nations, including to manage the risk that AI systems pose."* — United States Executive Order 14110 [1]

> *"All stakeholders, including industry, academia, civil society and standardisation organisations, are encouraged to take into account, as appropriate, the ethical principles for the development of voluntary best practices and standards."* — European Union AI Act [2]

This reflects a broader strategy of governance-by-consensus, based on reputational pressure and shared norms. The United Stated Biden-Harris Administration secured voluntary commitments from seven leading AI companies explicitly referencing watermarking to lead safe and transparent AI [31]. An additional group of eight companies later joined these commitments [30].

However, implementation remains fragmented, with many deployments being firm-specific and lacking interoperability. Few are publicly auditable or developed through shared infrastructure. In the absence of enforceable standards or independent oversight, these commitments risk enabling symbolic compliance, signaling safety while delivering little practical governance capability.

Voluntary alignment also depends on political continuity. In 2025, the United States Trump-Vance Administration rescinded Executive Order 14110, directing agencies to "suspend, revise, or rescind" related initiatives [32]. The administration's stated focus on technological competitiveness leaves the governance status of watermarking ambiguous, casting uncertainty over the future of watermarking as a governance tool. By contrast, China's Deep Synthesis Provisions[8] offer a more directive model, mandating compliance and audits, but even in such regimes, enforcement is uneven and robustness is not guaranteed.

While voluntary commitments may promote cooperation, they offer a fragile foundation for governance. When watermarking imposes costs or strategic risks, firms are unlikely to sustain alignment in the absence of legal or institutional compulsion.

3 Why the Gap Persists: Industry Incentives

Despite growing regulatory attention, industry watermarking implementations remain fragmented. This is not simply a technical lag but reflects a deeper misalignment between governance objectives and industry incentives. Without enforceable standards, firms have strong incentives to pursue symbolic or minimal compliance [3].

First, watermarking offers reputational value. Companies can signal alignment with governance goals by announcing watermarking initiatives, often without disclosing technical details or enabling verification. This satisfies public demands at low cost, while sidestepping the risks of committing to truly robust solutions.

Second, robust watermarking entails economic and strategic risks. If detection tools are open source, they can be exploited by modifying outputs to evade detection or crafting content that falsely appears watermarked. Conversely, restricted-access to detection capabilities would require companies to develop secure infrastructure to manage watermarking keys, detection thresholds, and access control via cryptographic or trust-based frameworks. These trade-offs make firms hesitant to invest in public, auditable watermarking systems that support third-party verification.

Third, companies face disincentives to move unilaterally. Without coordination, early adopters risk losing users if watermarking is perceived as restrictive, such as for moderation or IP enforcement, especially if competitors offer unmarked alternatives. In a rapidly evolving market, the incentive to retain users outweighs alignment with long-term governance objectives.

Finally, regulatory uncertainty undermines long-term planning, further amplifying industry hesitation. As seen with the revocation of Executive Order 14110 [32], shifting political priorities can quickly alter the expected policy landscape. In such an environment, companies are hesitant to invest in watermarking infrastructure that may soon be obsolete.

Together, these forces collectively encourage minimal, firm-specific, and performative watermarking environments, an outcome misaligned with the goals of effective AI oversight.

4 A Three-Layer Approach to Enforceable Watermarking

To bridge the gap between policy ambition and technical feasibility, we propose a three-layer framework: (I) technical requirements, (II) audit infrastructure, and (III) policy enforcement. Each layer targets a key weakness in current watermarking practices and together, they establish the groundwork for end-to-end accountability. Figure 1 summarizes the framework.

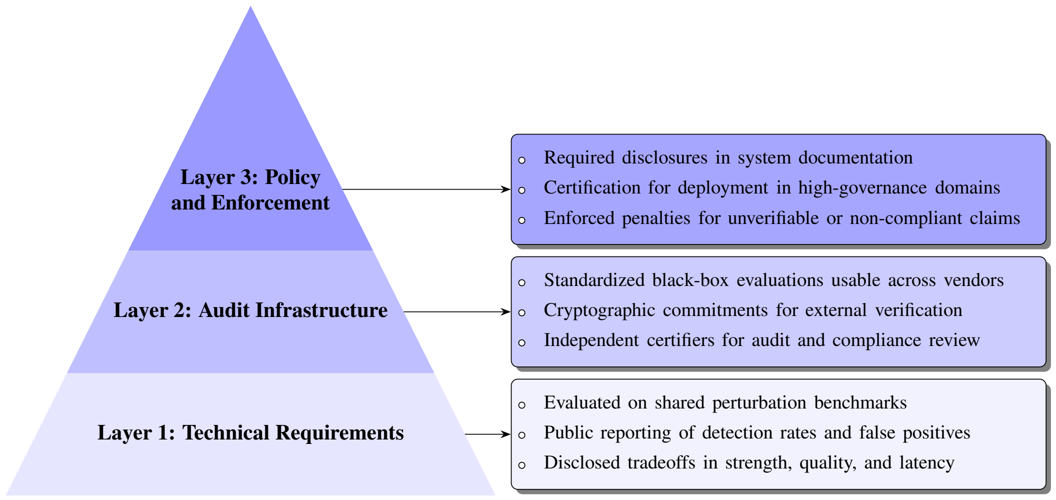
Layer 3: Policy and Enforcement
Required disclosures in system documentation
Certification for deployment in high-governance domains
Enforced penalties for unverifiable or non-compliant claims
Layer 2: Audit Infrastructure
Standardized black-box evaluations usable across vendors
Cryptographic commitments for external verification
Independent certifiers for audit and compliance review
Layer 1: Technical Requirements
Evaluated on shared perturbation benchmarks
Public reporting of detection rates and false positives
Disclosed tradeoffs in strength, quality, and latency

Figure 1: **A three-layer framework for enforceable watermarking.** Each layer represents a distinct governance function: technical guarantees, independent auditability, and regulatory enforcement. Arrows point to concrete mechanisms that instantiate the requirements at each level, linking system design to policy accountability.

4.1 Technical Requirements

Effective governance depends on watermarking systems that are technically sound and externally testable. At a minimum, watermarking methods should demonstrate robustness to both benign transformations and adversarial modifications designed to evade detection. Systems must report standardized performance metrics, including detection accuracy, false positive rates, and robustness under defined perturbation sets. Crucially, these guarantees should not require full transparency of proprietary model internals but must expose interfaces and artifacts, such as detectors, keys, or hashes, that enable reproducible evaluation by third parties. This layer establishes the minimum technical properties needed to claim that watermarking is doing more than signaling intent.

Watermarking spans multiple generative modalities, including text, image, and audio [34]. Each modality requires distinct robustness evaluations to assess watermark persistence under realistic conditions. Robustness refers to a watermark's ability to withstand both benign transformations (e.g., compression, summarization) and adversarial modifications (e.g., targeted perturbations designed to evade detection). Layer 1 calls for unified, reproducible technical standards to evaluate watermark durability against such transformations.

Each modality exhibits specific transformations that risk watermark integrity:

- Text: paraphrasing, spelling alterations, lexical substitutions, summarization, and machine translation [17].
- Image: lossy compression (e.g., JPEG), geometric manipulations (cropping, rotation, scaling), color shifts, and style transfer [35].
- Audio: background noise injection, pitch shifts, re-encoding artifacts, and time-stretching [10].

A watermarking scheme must be evaluated against these transformations not only for robustness, but also for recoverability without false attribution.

We suggest the creation of a shared benchmarking library, specifically tailored for generative watermarking evaluation. Such a library would include suites of transformation pipelines standardized across modalities and automated reporting of scores. Performance metrics should include, at minimum, detection accuracy under perturbation, false positive and false negative rates, and area under the ROC curve (AUC). Where applicable, modality-specific metrics should also be reported.

This infrastructure would be centrally maintained and periodically updated to reflect new attack strategies and model capabilities. In practice, this would enable both independent researchers and regulatory agencies to evaluate watermarking claims reproducibly across diverse systems. Systems must also document tradeoffs introduced by watermarking, such as reductions in model output quality, inference latency increases, or decreased robustness to noise so users and auditors understand the operational impact of embedding techniques. To stay ahead of emerging threats, the library should also support contributions from red-teaming initiatives and adversarial research challenges, ensuring that benchmark suites evolve alongside attack capabilities [4, 7].

While Executive Order 14110 [1] identified the National Institute of Standards and Technology (NIST) as a natural candidate for benchmarking watermarking technologies, the pace of AI system deployment exceeds NIST's current update cycles. We recommend that NIST serve as the registry and certifier, while operational development of the library should be delegated to a more agile, community-driven consortium. This hybrid structure would combine institutional legitimacy with rapid iteration and responsiveness to evolving technical landscapes.

To institutionalize this layer, any system claiming compliance must submit its watermarking pipeline to this library's API for evaluation. Recognizing potential concerns over intellectual property or security, alternative compliance paths such as exposing a detection API or providing zero-knowledge proofs could be considered, provided they allow equivalent third-party evaluation. Outputs will be scored on standardized perturbation suites, with results logged in a public registry. This shifts evaluation from self-reported metrics to verifiable compliance with shared robustness expectations. Such a system ensures that claims of watermark durability are not merely assertions, but demonstrable properties grounded in reproducible tests. These reproducible evaluations also form the empirical basis for audit infrastructure in Layer 2, enabling certifiers to assess compliance without needing privileged access.

4.2 Audit Infrastructure

Robust watermarking requires more than technical performance as it must be verifiable through independent and reproducible evaluation. This layer establishes the infrastructure for third-party audits, including standardized black-box testing protocols, shared testbeds, and recognized certifying bodies. Audits should reflect real-world use conditions rather than idealized lab settings and must work across different vendors without requiring access to proprietary internals. Systems should expose externally testable behaviors or cryptographic commitments that allow verification without reverse engineering. By enabling interoperable, independent testing, this layer transforms watermarking from internal assurance into a publicly accountable mechanism.

To verify that deployed watermarking schemes meet the minimal technical requirements outlined in Layer 1, robust audit systems must be established. However, watermarking must remain resilient to perturbations while concealing key detection parameters. Full public access to a detection system would enable adversaries to remove or forge watermarks, undermining the scheme's integrity. Consequently, audit systems must operate in a black-box setting where evaluators do not access model internals or watermarking code, but instead submit test content and observe binary or probabilistic detection outcomes. This approach introduces challenges since black-box detectors are vulnerable to repeated querying attacks, where adversaries iteratively probe a system to learn its boundaries [23]. To mitigate this, detection interfaces must implement access controls, either via restrained queries or containerized deployments [27]. These include vendor-specific configurations such as closed-source APIs or on-premise audit tools, depending on the operational context or vendor policy.

Watermarking deployments should also support cryptographic commitments that allow external verification without revealing sensitive internals. This can involve traditional hash-based attestations (e.g., SHA-256 commitments to watermark parameters) [22] or emerging cryptographic techniques such as zero-knowledge proofs [24]. For example, a system might prove to an auditor that it conforms to a certified detection threshold, without disclosing the underlying keys or watermark design.

A central question concerns auditor identity and governance. Potential certifiers include NIST-accredited laboratories, third-party security firms, or consortia established for compliance auditing [5]. Certification levels may include baseline compliance for systems meeting minimum robustness, interoperability certification for systems supporting cross-vendor detection, or adversarial resilience certification for schemes robust under threat-model testing. Audit cadence is equally critical. Evaluations should occur: (i) pre-deployment, to certify governance readiness, (ii) periodically post-deployment, to catch degradation or drift, and (iii) post-incident, in response to adversarial exploits or system failures. As adversaries evolve, the audit layer must ensure that claims of durability and detectability remain valid over time. Auditors must be able to reproduce robustness metrics defined in Layer 1 using only externally exposed interfaces. This includes performance on benchmark perturbation sets, detection behavior under adversarial conditions, and public system documentation. Certified systems should be logged in a public registry indicating audit status and any revocations or compliance failures. While audit processes introduce operational overhead, they establish trust by linking technical claims to external validation. Layer 2 builds directly on the technical guarantees of Layer 1, ensuring they are externally verified, reproducible, and trustworthy in deployment contexts. A prototype evaluation scorecard illustrating these technical benchmarks and audit criteria is shown in Table 2.

4.3 Policy and Enforcement

The final layer ensures that technical and audit standards translate into real-world accountability. Without legal mandates and institutional enforcement, even robust and verifiable watermarking systems may go unused or implemented inconsistently. These mechanisms should be tied to existing governance processes, such as model registration or domain-specific safety requirements. Public disclosures, for instance, in model cards, help standardize expectations and support auditability. Certification should be required for deployment in high-risk contexts, and penalties should apply for unverifiable or misleading claims. This layer connects the technical and audit layers to institutional incentives, closing the loop from design to deployment.

To ensure that technical requirements and audit mechanisms translate into industry commitment, enforceable legal and regulatory consequences must be in place. Without binding consequences, the outputs of Layers 1 and 2 risk being symbolic, allowing firms to claim compliance without delivering accountability. Layer 3 operationalizes these technical and audit layers by anchoring them in policy frameworks that mandate compliance and impose consequences for deviation.

Certification based on audit results should be required for deployment in high-governance contexts (e.g., elections, education, public interfaces). Systems that fail certification would face deployment restrictions or public disclosure of non-compliance. In tandem, mandated disclosures in system documentation and model cards must include audit status, robustness guarantees, and known failure modes [18]. These disclosures provide transparency while enabling downstream accountability. Non-compliant systems failing to meet audit standards or refusing audit participation should face graduated enforcement actions. These may include fines, removal of deployment licenses for regulated sectors, or public listing in non-compliance registries. For firms, such penalties also carry reputational risk, further incentivizing alignment with certification pathways.

Enforcement authority must be clearly defined. Governance bodies such as NIST (United States) [1], designated regulators under the European Union AI Act [2], or China's Cyberspace Administration [8] can serve as enforcement agents within their jurisdictions. However, because AI systems operate across borders, enforcement must also address international fragmentation. A model deployed in one jurisdiction, producing outputs accessible in another, would make alignment across regimes critical for consistent governance.

Finally, policy enforcement must evolve alongside the underlying technologies. Just as audit protocols adapt to new adversarial techniques, enforcement mechanisms must include provisions for periodic policy review and revision. Feedback loops from auditors, researchers, and affected stakeholders can help keep regulatory frameworks responsive and legitimate. By linking technical performance and audit compliance to real-world consequences, Layer 3 closes the loop from design to deployment. It ensures that watermarking systems are not merely well-engineered, but meaningfully accountable in practice.

5 Alternative Views

While we argue for enforceable watermarking as a critical step toward meaningful AI governance to transform it from a symbolic gesture into a mandatory mechanism, it is important to acknowledge opposing perspectives that highlight real implementation challenges.

Our proposed three-layer framework, though necessary for governance, poses substantial practical hurdles. The technical complexity alone requires the development of robust watermarking schemes across modalities, supported by standardized benchmarks and test protocols. Beyond technical design, the framework calls for the creation of new institutions to oversee audit infrastructure and certification processes. These demands would foreseeably slow down the current rapid pace of generative AI innovation, particularly for smaller firms and open-source communities.

In addition to technical challenges, the fragmented landscape of both AI development and global regulation limits the interoperability and enforceability of watermarking systems. Proprietary models often adopt incompatible watermarking implementations, while open-source forks can bypass governance requirements entirely. On the regulatory side, jurisdictions such as the United States, European Union, and People's Republic of China are developing divergent standards, ranging from voluntary commitments to strict mandates, resulting in conflicting requirements. A model compliant in one region may not meet the standards of another, undermining global accountability efforts.

Amid this fragmentation, some would argue that voluntary compliance and soft norms offer a more flexible and innovation-friendly alternative to rigid enforcement. As noted in Section 2.3, shifts in United Stated policy have demonstrated this volatility while one administration secured voluntary commitments from fifteen major AI firms, a subsequent administration rescinded those governance efforts. For some, this approach appears more feasible and politically viable than establishing binding standards. However, we contend that voluntary commitments alone are insufficient. In the absence of enforceable requirements and independent audits, companies are more likely to implement minimal or symbolic watermarking measures that fall short of supporting meaningful oversight.

6 Conclusion

Watermarking is rapidly becoming a pillar of AI governance, but without enforceable standards and verifiable implementation, it risks becoming a symbol of oversight rather than a mechanism of accountability. We argue that robustness, verifiability, and auditability must be built into watermarking from the ground up and not added as compliance afterthoughts. Realizing this vision requires a structural shift for clear technical baselines, independent audit infrastructure, and regulatory enforcement that binds them together. Achieving durable accountability will depend on sustained collaboration between policymakers, industry actors, and the technical research community.

Table 2: Evaluation Scorecard for Layer 1 Watermarking Requirements

Category	Metric	Score Range	Comments / Benchmarks
Robustness (Text) [33]	Paraphrasing	0–5	5 = survives 90%+ of semantic paraphrasing tools
	Token substitution	0–5	5 = <5% detection drop with synonym swaps
	Translation	0–5	5 = survives round-trip machine translation in 3+ languages
	…	…	…
Robustness (Image) [6]	JPEG compression	0–5	5 = survives down to 20% quality with ≥95% detectability
	Cropping & rotation	0–5	5 = survives ±10° rotation and 15% central crop
	Color/style shift	0–5	5 = persists across style transfer and color remapping
	…	…	…
Robustness (Audio) [29]	Noise injection	0–5	5 = survives pink/white noise @ 20 dB SNR
	Time-stretching	0–5	5 = survives ±5% tempo shift with detectable signal
	…	…	…
Detection Quality	False Positive Rate	0–5	5 = <0.1% FPR on curated human-generated corpora
	False Negative Rate	0–5	5 = <1% FN under defined threat model
	…	…	…
Auditability	Third-party reproducibility	0–5	5 = Public test suite, verified by 2+ independent labs
	Detection API/tool availability	0–5	5 = Public API or SDK with access controls
	…	…	…
Modality Coverage	Text, Image, Audio support	0–5	1 point per supported modality
Security Disclosure	Known failure modes documented	0–5	5 = Threat model or model card made public
Policy Alignment	Standards-based declaration	Yes/No	Must align with NIST or equivalent regulatory framework
	…	…	…
…	…	…	…

References

[1] Executive order 14110 of october 30, 2023: Safe, secure, and trustworthy development and use of artificial intelligence, November 2023. Federal Register, Vol. 88, No. 210, pp. 75191–75208.

[2] Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending certain union legislative acts (artificial intelligence act), July 2024. Official Journal of the European Union, L 2024/1689, 12 July 2024.

[3] S. Aaronson. Ai safety and watermarking. Talk at Simons Institute, UC Berkeley, 2024.

[4] L. Ahmad, S. Agarwal, M. Lampe, and P. Mishkin. Openai's approach to external red teaming for ai models and systems. *arXiv preprint arXiv:2503.16431*, 2025.

[5] N. AI. Artificial intelligence risk management framework (ai rmf 1.0). *URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai*, pages 100–1, 2023.

[6] B. An, M. Ding, T. Rabbani, A. Agrawal, Y. Xu, C. Deng, S. Zhu, A. Mohamed, Y. Wen, T. Goldstein, et al. Waves: Benchmarking the robustness of image watermarks. *arXiv preprint arXiv:2401.08573*, 2024.

[7] Anthropic. Progress from our frontier red team, 2025. Accessed: May 12, 2025.

[8] CAC. Provisions on the administration of deep synthesis internet information services, November 2022. Effective January 10, 2023.

[9] California State Legislature. California ai transparency act, sb 942, chapter 291, September 2024. Approved by the Governor on September 19, 2024; effective January 1, 2026.

[10] G. Chen, Y. Wu, S. Liu, T. Liu, X. Du, and F. Wei. Wavmark: Watermarking for audio generation. *arXiv preprint arXiv:2308.12770*, 2023.

[11] S. Dathathri, A. See, S. Ghaisas, P.-S. Huang, R. McAdam, J. Welbl, V. Bachani, A. Kaskasoli, R. Stanforth, T. Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

[12] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023.

[13] S. A. Fisher, J. W. Howard, and B. Kira. Moderating synthetic content: The challenge of generative ai. *Philosophy & Technology*, 37(4):133, 2024.

[14] Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[15] Google DeepMind. Synthid, 2023. Accessed: 2025-05-09.

[16] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[17] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR, 2023.

[18] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[19] A. Nemecek, Y. Jiang, and E. Ayday. Topic-based watermarks for llm-generated text. *arXiv preprint arXiv:2404.02138*, 2024.

[20] OpenAI. New ai classifier for indicating ai-written text, January 2023. Tool discontinued as of July 20, 2023 due to low accuracy.

[21] L. Pan, A. Liu, Z. He, Z. Gao, X. Zhao, Y. Lu, B. Zhou, S. Liu, X. Hu, L. Wen, et al. Markllm: An open-source toolkit for llm watermarking. *arXiv preprint arXiv:2405.10051*, 2024.

[22] F. Pub. Secure hash standard (shs). *Fips pub*, 180(4):180–4, 2012.

[23] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

[24] E. B. Sasson, A. Chiesa, C. Garman, M. Green, I. Miers, E. Tromer, and M. Virza. Zerocash: Decentralized anonymous payments from bitcoin. In *2014 IEEE symposium on security and privacy*, pages 459–474. IEEE, 2014.

[25] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

[26] D. H. Spennemann. Delving into: the quantification of ai-generated content on the internet (synthetic data). *arXiv preprint arXiv:2504.08755*, 2025.

[27] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pages 601–618, 2016.

[28] U.S. Senate. Artificial intelligence research, innovation, and accountability act of 2023, s. 3312, 118th congress, November 2023. Introduced by Sen. John Thune on November 15, 2023; reported with amendment on December 18, 2024.

[29] Y. Wen, A. Innuganti, A. B. Ramos, H. Guo, and Q. Yan. Sok: How robust is audio watermarking in generative ai models? *arXiv preprint arXiv:2503.19176*, 2025.

[30] White House. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI, September 2023. Accessed: 2025-05-09.

[31] White House. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI, July 2023. Accessed: 2025-05-09.

[32] White House. Executive Order 14179: Removing Barriers to American Leadership in Artificial Intelligence, January 2025. Accessed: 2025-05-09.

[33] Z. Yang, G. Zhao, and H. Wu. Watermarking for large language models: A survey. *Mathematics*, 13(9), 2025.

[34] X. Zhao, S. Gunn, M. Christ, J. Fairoze, A. Fabrega, N. Carlini, S. Garg, S. Hong, M. Nasr, F. Tramer, S. Jha, L. Li, Y.-X. Wang, and D. Song. Sok: Watermarking for ai-generated content. *arXiv preprint arXiv:2411.18479*, 2024.

[35] X. Zhao, K. Zhang, Z. Su, S. Vasan, I. Grishchenko, C. Kruegel, G. Vigna, Y.-X. Wang, and L. Li. Invisible image watermarks are provably removable using generative ai. *Advances in Neural Information Processing Systems*, 37:8643–8672, 2024.

A Prompts and Example Outputs for Watermark Robustness Evaluation

To supplement the evaluation results shown in Table 1, Table 3 lists the full prompt texts and transformation instructions used during robustness testing. Additionally, Table 4 presents complete original, paraphrased, and summarized outputs for one of the five selected prompts. These examples illustrate how minor but natural edits can significantly affect detection results.

Table 3: Evaluation prompts and transformation instructions used in robustness testing. Each output was paraphrased and summarized using GPT-4o with the listed instructions.

Prompt ID	**Prompt Text**
Prompt_1	Summarize the causes and impacts of climate change.
Prompt_2	Write a product review for a product you recently used.
Prompt_3	Draft an email apologizing for a delayed response.
Prompt_4	Explain quantum computing to a high school student.
Prompt_5	Tell a short story about a robot learning emotions.
Transformation	**Instruction Text**
Paraphrase	Paraphrase the following text while preserving its original meaning as closely as possible.
Summarize	Summarize the following text into 2–3 sentences, keeping the most important information.

Table 4: Full outputs for original, paraphrased, and summarized versions of Prompt$_1$. The original text includes markdown-style formatting (e.g., ##, **) (i.e., from ChatGPT's copy function).

Prompt$_1$: Summarize the causes and impacts of climate change.	
Original	## Climate Change: Causes & Impacts **Causes:** * **Greenhouse Gas Emissions:** Primarily from burning fossil fuels (coal, oil, gas), deforestation, and industrial processes. These gases trap heat in the atmosphere, leading to a warming effect. * **Human Activities:** Agriculture, transportation, and energy production significantly contribute to greenhouse gas emissions. * **Deforestation:** Trees absorb CO2, so cutting them down reduces the planet's ability to absorb these gases. **Impacts:** * **Rising Global Temperatures:** Leading to more frequent and intense heatwaves, droughts, and wildfires. * **Sea Level Rise:** Melting glaciers and thermal expansion of water threaten coastal communities and ecosystems. * **Extreme Weather Events:** Increased frequency and intensity of storms, floods, and droughts, causing widespread damage and displacement. * **Ocean Acidification:** Absorption of CO2 by oceans makes them more acidic, harming marine life and ecosystems.
Paraphrased	## Climate Change: Causes and Effects **Causes:** * **Greenhouse Gas Emissions:** Mostly result from burning fossil fuels such as coal, oil, and natural gas, as well as from industrial activities and deforestation. These gases trap heat in the Earth's atmosphere, causing it to warm. * **Human Activities:** Sectors like agriculture, transportation, and energy generation are major contributors to the release of greenhouse gases. * **Deforestation:** Trees play a key role in absorbing carbon dioxide. When forests are cleared, this natural absorption capacity is diminished, increasing atmospheric CO2 levels. **Effects:** * **Increasing Global Temperatures:** This leads to more intense and frequent heatwaves, wildfires, and periods of drought. * **Rising Sea Levels:** Caused by melting ice and the expansion of warmer ocean water, posing risks to coastal regions and habitats. * **More Severe Weather Events:** Storms, floods, and droughts are becoming more common and intense, resulting in significant destruction and displacement of populations. * **Ocean Acidification:** Oceans absorb large amounts of carbon dioxide, which lowers their pH and negatively affects marine organisms and ecosystems.
Summarized	Climate change is primarily driven by human activities such as burning fossil fuels, deforestation, and industrial processes, which release greenhouse gases that trap heat in the atmosphere. Its impacts include rising global temperatures, more extreme weather events, sea level rise, and ocean acidification, all of which pose serious threats to ecosystems and human communities.