Fragile by Design: Formalizing Watermarking Tradeoffs via Paraphrasing

Anonymous Authors[1]

Abstract

Verification is a cornerstone of AI governance, enabling auditability, attribution, and accountability in AI-generated content. As generative models proliferate, watermarking has emerged as a leading verification strategy. However, state-of-the-art paraphrasing methods pose a serious threat: they can erase watermarks without altering the meaning of the generated output. We model watermarking under paraphrasing as an adversarial game and prove a no-go theorem: under idealized conditions, no watermark can be both robust to paraphrasing and imperceptible. To navigate this tension, we propose the ε–δ framework that quantifies the trade-off between robustness (ε) and semantic distortion of the underlying text (δ). Our findings highlight a key asymmetry: removing a watermark is often easier than embedding one that survives.

1. Introduction

As large language models (LLMs) increasingly power generative systems, questions of authorship, auditability, and accountability become more urgent. In safety-critical and socially impactful applications, technical mechanisms for tracing provenance are essential for governance. Watermarking, which embeds algorithmically-detectable statistical signals in generated text, has emerged as a significant strategy to support attribution, flag misuse, and enable post hoc audit (Kirchenbauer et al., 2023).

However, watermarking is fragile. The primary adversary is not cryptanalysis but paraphrasing, the automated tools that rewrite text while preserving its meaning. A single pass through a modern paraphraser can erase a watermark's statistical signature without altering the semantics (Ren et al., 2023; Hou et al., 2024). This problem is amplified by *iterative paraphrasing*, where multiple rewriting stages slowly degrade the watermark until detection becomes impossible (Zhang et al., 2023; Cohen et al., 2024). We therefore ask: **is robust, imperceptible watermarking fundamentally possible when adversaries can paraphrase?**

Despite the awareness of the trade-off between robustness and imperceptibility, this tension has not been made precise. To fill this gap, we articulate this trade-off explicitly through paraphrasing. By treating paraphrasing as an adversarial process that preserves semantics, we expose the incompatibility between watermarking robustness and invisibility.

To do so, we propose a framework for watermarking under meaning-preserving paraphrasers. By modeling semantic equivalence and paraphrasing as operations over sentence sets, we derive an *impossibility result*: when the paraphraser can access the full set of semantically equivalent rewritings, watermark detection breaks down completely. No watermarking scheme can separate watermarked from unmarked texts without semantic distortion.

To move beyond this deadlock, we introduce the ε,δ **watermarking framework** that captures the trade-off between robustness (ε) and semantic distortion (δ). Here, ε measures how well a watermark resists paraphrasing, while δ quantifies the resulting semantic deviation. This trade-off is not just theoretical; we empirically map it across several state-of-the-art watermarking schemes, revealing distinct robustness-fidelity profiles.

2. Problem Setup: The Paraphrasing Game

We represent text as a sequence of sentences drawn from a language. To account for both syntactic and semantic watermarking, we move beyond individual tokens and instead model watermarking at the level of coherent units of meaning. We focus on sentences as the minimal context-bearing units, though our analysis naturally extends to larger structures such as paragraphs.

Let $\mathcal{L}$ be the set of all possible sentences, and let $\mathcal{T} = \mathcal{L}^n$ denote the space of texts of fixed length n. For any text $S = (s_1, s_2, \ldots, s_n) \in \mathcal{T}$, we define semantic equivalence as a binary relation $\equiv \subseteq \mathcal{T} \times \mathcal{T}$, where $X \equiv Y$ if and only if X and Y convey the same meaning. This equivalence may be judged by human annotators or approximated algorithmically via semantic similarity models.

For a given position i in a text S, we define $\mathcal{M}_i(S)$, the set of sentences comprising semantically valid substitutions:

$$\mathcal{M}_i(S) = \{s \in \mathcal{L} \mid (s_1, \ldots, s_{i-1}, s, s_{i+1}, \ldots, s_n) \equiv S\}$$

If two texts X and Y are semantically equivalent, it follows

that $\mathcal{M}_i(X) = \mathcal{M}_i(Y)$ for all i.

Definition 2.1. A *paraphraser* P assigns to each text $S \in \mathcal{T}$ and position $i \in \{1, \ldots, n\}$ a set $P_i(S) \subseteq \mathcal{M}_i(S)$ of permissible replacements. The full set of paraphrases produced by P is the Cartesian product:

$$\mathcal{P}(S) = \prod_{i=1}^{n} P_i(S)$$

We say P is *perfect* if $P_i(S) = \mathcal{M}_i(S)$ for all S and i, meaning that it explores the full semantic equivalence class of S. Otherwise, P is *imperfect*.

The Paraphrasing Game. Watermarking plays out as a game between two players: the watermarking scheme and the paraphraser. The scheme embeds a signal into a text Y that is semantically equivalent to an original text X. Acting as the adversary, the paraphraser attempts to rewrite Y and potentially X without changing the meaning. The scheme aims to design a detector D that distinguishes paraphrases of Y from those of X, even after multiple rounds of rewriting.

Let $X = (x_1, \ldots, x_n)$ be an original text and $Y = (y_1, \ldots, y_n)$ its watermarked counterpart, where $X \equiv Y$ and each $x_i, y_i \in \mathcal{L}$.

Definition 2.2. A *watermarking scheme* under a paraphraser P consists of (1) a watermarked text $Y \in \mathcal{T}$ such that $X \equiv Y$, and (2) a detector $D : \mathcal{T} \to \{0, 1\}$ satisfying

(1) For all $W \in \mathcal{P}(Y)$, $D(W) = 1$ (true positive);

(2) For all $Z \in \mathcal{P}(X)$, $D(Z) = 0$ (true negative). The scheme is *robust* against P if such a D exists.

Why Robustness Is Fragile. A robust detector must reliably fire on any paraphrase of the watermarked text Y, and never fire on any paraphrase of the original text X. But this condition is unstable under composition: if a paraphrased version $W \in \mathcal{P}(Y)$ is further paraphrased into $W' \in \mathcal{P}(X)$ by a different paraphraser P', the watermark vanishes.

This leads to a fundamental insight: to ensure robustness under *any* sequence of meaning-preserving rewrites, the detector must separate *entire paraphrase sets* $\mathcal{P}(Y)$ and $\mathcal{P}(X)$. This sets the stage for our upcoming impossibility result.

3. Robustness Under Perfect Paraphrasers

We now consider the most adversarial case: a *perfect paraphraser* that exhaustively explores the entire semantic equivalence class of a given text. This means $P_i(S) = \mathcal{M}_i(S)$ for all positions i and all texts S. In this setting, any sentence replacement that preserves meaning is accessible to the paraphraser. As we show below, this level of semantic flexibility eliminates any hope of robust watermarking.

Theorem 3.1. *Let P be a perfect paraphraser. Then for any pair of semantically equivalent texts $X \equiv Y$, we have $\mathcal{P}(X) = \mathcal{P}(Y)$. Consequently, no robust watermarking scheme can exist against P.*

Proof. By definition of semantic equivalence, $X \equiv Y$ implies that $\mathcal{M}_i(X) = \mathcal{M}_i(Y)$ for all i. Since P is perfect:

$$P_i(X) = \mathcal{M}_i(X) = \mathcal{M}_i(Y) = P_i(Y) \quad \forall i$$

Taking Cartesian products:

$$\mathcal{P}(X) = \prod_{i=1}^{n} P_i(X) = \prod_{i=1}^{n} P_i(Y) = \mathcal{P}(Y)$$

So $\mathcal{P}(X) = \mathcal{P}(Y)$. But this contradicts the goal of any watermarking detector D: to assign $D(W) = 1$ for all $W \in \mathcal{P}(Y)$ and $D(Z) = 0$ for all $Z \in \mathcal{P}(X)$. Since these two sets are equal, no function D can satisfy both constraints without contradiction. □

This result highlights a fundamental limitation. When the paraphraser has unrestricted access to the full semantic equivalence class

$$[X] = \{T \in \mathcal{T} \mid T \equiv X\},$$

any watermark embedded in $Y \equiv X$ becomes impossible to detect, since both texts generate the exact same paraphrase space. Thus, *any watermark robust to perfect paraphrasing must necessarily alter meaning*. But doing so violates the imperceptibility principle that motivates watermarking in the first place. This impossibility theorem sets a hard limit: Semantic-preserving robustness and invisibility cannot coexist in the presence of ideal paraphrasers.

4. Robustness Under Imperfect Paraphrasers

The impossibility of robust watermarking under perfect paraphrasers stems from their ability to fully traverse the semantic equivalence class of a text. But in practice, paraphrasers operate under architectural constraints, training data biases, and decoding limits. Let us therefore consider an *imperfect paraphraser* P such that $P_i(S) \subsetneq \mathcal{M}_i(S)$ for some i and some S. Unlike the perfect case, it is now possible for the paraphrase sets $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ to be *disjoint*. When this occurs, a watermark detector can perfectly distinguish paraphrases of Y from those of X.

Theorem 4.1. *Let P be an imperfect paraphraser. Then $\mathcal{P}(X) \cap \mathcal{P}(Y) = \emptyset$ if and only if there exists an index $i \in \{1, \ldots, n\}$ such that $P_i(X) \cap P_i(Y) = \emptyset$. When this holds, a robust watermarking scheme against P exists.*

Proof. Observe that:

$$\mathcal{P}(X) = \prod_{i=1}^{n} P_i(X), \quad \mathcal{P}(Y) = \prod_{i=1}^{n} P_i(Y)$$

and thus their intersection becomes:

$$\mathcal{P}(X) \cap \mathcal{P}(Y) = \prod_{i=1}^{n} (P_i(X) \cap P_i(Y))$$

This Cartesian product is non-empty if and only if $P_i(X) \cap P_i(Y) \neq \emptyset$ for all i. Therefore, $\mathcal{P}(X) \cap \mathcal{P}(Y) = \emptyset$ if and only if $P_i(X) \cap P_i(Y) = \emptyset$ for some i. If such an i exists, then any text $Z \in \mathcal{P}(X)$ must have $z_i \in P_i(X)$, and thus $z_i \notin P_i(Y)$. So $Z \notin \mathcal{P}(Y)$. The same logic holds symmetrically. Hence, $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ are disjoint.

We can then define a detector $D : \mathcal{T} \to \{0, 1\}$ by:

$$D(T) = \begin{cases} 1 & \text{if } t_i \in P_i(Y) \\ 0 & \text{if } t_i \in P_i(X) \\ 0 & \text{otherwise} \end{cases}$$

Because $P_i(X) \cap P_i(Y) = \emptyset$, this detector makes no errors on any paraphrases of X or Y. Thus, the watermark is robust against P. □

This theorem shows that robustness is *theoretically* possible when the paraphraser's choices for X and Y differ enough at even a single position. But this robustness is inherently fragile. It depends entirely on the specific subsets $P_i(X)$ and $P_i(Y)$ chosen by the paraphraser. If an adversary switches to a different paraphraser P' where the corresponding replacement sets overlap (i.e., $P'_i(X) \cap P'_i(Y) \neq \emptyset$), then $\mathcal{P}'(X)$ and $\mathcal{P}'(Y)$ may again intersect, and the watermark becomes undetectable.

Robustness Is Not Composable. To understand the fragililty of robustness under imperfect paraphrasers, suppose the set of meaning-preserving replacements at position i is fixed for both texts:

$$\mathcal{M}_i(X) = \mathcal{M}_i(Y) = \{s_1, s_2, s_3\}$$

Now consider a paraphraser P that selects disjoint subsets: $P_i(X) = \{s_1\}$ and $P_i(Y) = \{s_2\}$. According to Theorem 4.1, this yields disjoint paraphrase sets $\mathcal{P}(X) \cap \mathcal{P}(Y) = \emptyset$, and robustness is achievable. But suppose an adversary switches to a more permissive paraphraser P' with:

$$P'_i(X) = \{s_1, s_3\}, \quad P'_i(Y) = \{s_2, s_3\}$$

Now, the overlap $s_3 \in P'_i(X) \cap P'_i(Y)$ reintroduces ambiguity. Any paraphrased text containing s_3 at position i could belong to both $\mathcal{P}'(X)$ and $\mathcal{P}'(Y)$, breaking the detector's ability to reliably distinguish between them.

This illustrates a critical point: *robustness under imperfect paraphrasing is not stable under paraphraser composition or variation*. The watermark may survive one rewriting strategy but collapse under another. This motivates our next step: a framework that quantifies this trade-off rather than treating it as binary success or failure.

5. ε–δ Watermarks

The impossibility of robust watermarking under perfect paraphrasers, and the fragility of robustness under imperfect ones, point to a deeper tension: watermarking must trade off *semantic preservation* and *robust detection*. To make this trade-off explicit, we introduce the notion of an ε–δ watermark. This framework defines two continuous axes:

- ε quantifies *robustness*, how well a watermark survives paraphrasing;
- δ quantifies *semantic distortion*, how much the watermarked text deviates in meaning from the original.

No scheme can minimize both simultaneously. High robustness tends to require detectable perturbations; high fidelity makes the watermark easier to erase. The ε–δ framework allows us to measure and compare how watermarking schemes navigate this fundamental trade-off.

Let X be an original text and Y its watermarked counterpart, with $X \equiv Y$. We apply a paraphraser P to each text and collect N paraphrases:

$$\mathcal{X} = \{x_i\}_{i=1}^{N} = P(X), \quad \mathcal{Y} = \{y_j\}_{j=1}^{N} = P(Y)$$

Let $\gamma \in [0, 1]$ be a similarity threshold. For each pair $(x_i, y_j) \in \mathcal{X} \times \mathcal{Y}$, we compute the cosine distance between their embeddings:

$$d_{\cos}(x_i, y_j) = 1 - \frac{E(x_i) \cdot E(y_j)}{\|E(x_i)\| \, \|E(y_j)\|}$$

If this distance is below γ, the pair is deemed semantically overlapping. We define the *thresholded intersection set*:

$$\mathcal{X} \cap_\gamma \mathcal{Y} = \{(x_i, y_j) \mid d_{\cos}(x_i, y_j) < \gamma\}$$

The union size is given by:

$$|\mathcal{X} \cup \mathcal{Y}| = |\mathcal{X}| + |\mathcal{Y}| - |\mathcal{X} \cap_\gamma \mathcal{Y}|$$

We define the robustness score as the Jaccard distance between the paraphrase sets:

$$\varepsilon = 1 - \frac{|\mathcal{X} \cap_\gamma \mathcal{Y}|}{|\mathcal{X} \cup \mathcal{Y}|}$$

A higher ε indicates lower semantic overlap between paraphrases of the original and watermarked text, implying greater resilience to paraphrasing attacks.

To quantify the semantic distortion, δ, we restrict attention to the subset of paraphrase pairs $(x_i, y_j) \in \mathcal{X} \times \mathcal{Y}$ that fall within the semantic threshold γ. That is, we compute the average pairwise cosine distance only over the intersection set:

$$\mathcal{X} \cap_\gamma \mathcal{Y} = \{(x_i, y_j) \mid d_{\cos}(x_i, y_j) < \gamma\}$$

We then define:

$$\delta = \frac{1}{|\mathcal{X} \cap_\gamma \mathcal{Y}|} \sum_{(x_i, y_j) \in \mathcal{X} \cap_\gamma \mathcal{Y}} d_{\cos}(x_i, y_j)$$

This formulation captures the average semantic drift only among paraphrase pairs deemed close in meaning. A smaller δ implies that even within the overlapping region, the watermarked paraphrases remain semantically faithful to the originals.

The ε–δ framework makes the trade-off between robustness and semantic preservation explicit:

- **High ε, High δ:** The watermark is robust, but meaning has likely been altered.
- **Low ε, Low δ:** The watermark is imperceptible but fragile; it vanishes under paraphrasing.

In practice, watermarking schemes must balance these opposing forces. The goal is not to maximize both ε and δ, but to find a Pareto-optimal operating point.

6. Experiments

We evaluate the ε–δ trade-off across six representative publicly-available watermarking schemes: KGW (Kirchenbauer et al., 2023), which increases the probability of generation of selected tokens, UNIGRAM (Zhao et al., 2023), which is a robust extension of KGW, SWEET (Lee et al., 2023), designed for watermarking code, EWD (Lu et al., 2024), which watermarks high-entropy tokens to avoid garbling low-entropy sequences, UPV (Liu et al., 2023), which uses a public key for watermark detection as opposed to a private key for both generation and detection, and EXP (Kuditipudi et al., 2023) that works with sequences rather than individual tokens. These methods span a spectrum of design philosophies, from token-level perturbations to syntactic and semantic rewrites.

Setup. For each scheme, we select 50 input sentences sampled from C4 (Raffel et al., 2020), a large common-crawl based dataset, and generate corresponding watermarked outputs using the method's default configuration. To simulate a paraphrasing attack, we apply a state-of-the-art publicly available paraphraser, Parrot (Damodaran, 2021), to both the original and watermarked texts, generating 30 paraphrases for each sentence. This results in two sets of paraphrases, denoted by $\mathcal{X}$ (original) and $\mathcal{Y}$ (watermarked). We define ε as a robustness metric, computed as the Jaccard distance between $\mathcal{X}$ and $\mathcal{Y}$, where set intersection is determined by pairwise cosine distances between sentence embeddings falling below a threshold γ. The semantic distortion δ is defined as the average pairwise cosine distance between the elements of $\mathcal{X}$ and $\mathcal{Y}$ in the intersection set, quantifying the extent of semantic drift introduced by watermarking. Sentence embeddings are obtained using a pretrained BERT encoder $E(\cdot)$ (Devlin et al., 2019). By varying the similarity threshold γ, we obtain different values of ε to trace the robustness-distortion trade-off.

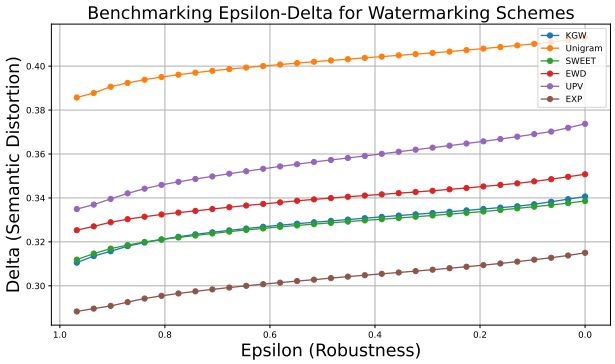

Figure 1. ε–δ trade-off curves for watermarking schemes.

Results. Figure 1 shows the ε–δ curves for the six watermarking schemes under increasing paraphrasing strength (left to right). As robustness (ε) increases, so does semantic distortion (δ), as expected. EXP dominates, with the lowest distortion across all robustness levels, which is surprising given its aggressive editing strategy. SWEET and KGW follow closely, with a balance between fidelity and robustness. In contrast, UNIGRAM performs worst: it has the highest semantic distortion ($\delta > 0.40$) even at low robustness. UPV and EWD lie in the mid-range: they show stable but elevated distortion as robustness increases. Notably, none of the methods achieves both low δ and high ε simultaneously, reinforcing our claim that watermarking lies on a fundamental trade-off curve. Our ε–δ framework makes these trade-offs explicit, offering a diagnostic tool for aligning watermarking designs with application-specific goals, be it resilience against adversaries or imperceptibility for benign attribution.

7. Conclusion

We formalized watermarking under paraphrasing as an adversarial game and proved that under perfect paraphrasers, robust and invisible watermarking cannot coexist. We then introduced the ε–δ watermarking framework to quantify the trade-off between robustness and semantic preservation. Empirical benchmarks on six watermarking schemes confirmed this trade-off, revealing that each method implicitly selects a point along the robustness–fidelity curve. We hope that our framework will be useful for evaluating new methods and setting appropriate expectations.

References

Or Cohen, Otto Eronen, Teemu Ernvall, Sharon Goldberg, Daniele Naldi, and Rafael Pass. Watermarking language models with adaptive adversaries. *IACR Cryptology ePrint Archive*, 2024:759, 2024.

Prithiviraj Damodaran. Parrot paraphraser: A paraphrasing toolkit for augmenting intent classification and natural language understanding tasks. https://github.com/PrithivirajDamodaran/Parrot_Paraphraser, 2021. Accessed: 2025-04-09.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186. Association for Computational Linguistics, 2019.

Yifan Hou, Canwen Xu, Renqian Zhang, Xin Wang, Lei Hou, Zhiyuan Liu, and Maosong Sun. Semstamp: Semantic invariant watermarking for language models. *arXiv preprint arXiv:2402.11399*, 2024.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR, 2023.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.

Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. Who wrote this code? watermarking for code generation. *arXiv preprint arXiv:2305.15060*, 2023.

Aiwei Liu, Leyi Pan, Xuming Hu, Shu'ang Li, Lijie Wen, Irwin King, and Philip S Yu. An unforgeable publicly verifiable watermark for large language models. *arXiv preprint arXiv:2307.16230*, 2023.

Yijian Lu, Aiwei Liu, Dianzhi Yu, Jingjing Li, and Irwin King. An entropy-based text watermarking detection method. *arXiv preprint arXiv:2403.13485*, 2024.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Jie Ren, Han Xu, Yiding Liu, Yingqian Cui, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. A robust semantics-based watermark for large language model against paraphrasing. *arXiv preprint arXiv:2311.08721*, 2023.

Hanlin Zhang, Benjamin L Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. *arXiv preprint arXiv:2311.04378*, 2023.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*, 2023.