More Haste, Less Speed: Weaker Single-Layer Watermark Improves Distortion-Free Watermark Ensembles

Ruibo Chen[1], **Yihan Wu**[1], **Xuehao Cui**[1], **Jingqi Zhang**[2], **Heng Huang**[1]
[1]University of Maryland, College Park, [2]National University of Singapore
{rbchen,ywu42,cedrus,heng}@umd.edu, zhangjingqi@u.nus.edu

arXiv:2602.11793v2 [cs.CR] 28 Aug 2026

Abstract

Watermarking has emerged as a crucial technique for detecting and attributing content generated by large language models. While recent advancements have utilized watermark ensembles to enhance robustness, prevailing methods typically prioritize maximizing the strength of the watermark at every individual layer. In this work, we identify a critical limitation in this "stronger-is-better" approach: strong watermarks significantly reduce the entropy of the token distribution, which paradoxically weakens the effectiveness of watermarking in subsequent layers. We theoretically and empirically show that detectability is bounded by entropy and that watermark ensembles induce a monotonic decrease in both entropy and the expected green-list ratio across layers. To address this inherent trade-off, we propose a general framework that utilizes weaker single-layer watermarks to preserve the entropy required for effective multi-layer ensembling. Empirical evaluations demonstrate that this counter-intuitive strategy mitigates signal decay and consistently outperforms strong baselines in both detectability and robustness.

1 Introduction

Large language models have demonstrated unprecedented capabilities in generating high-quality text, raising significant concerns regarding copyright infringement, misinformation, and the potential misuse of AI-generated content (Goldstein, 2023; Weidinger et al., 2021; Carlini et al., 2022; Chen et al., 2025b). To address these challenges, watermarking (Aaronson and Kirchner, 2022; Kirchenbauer et al., 2023; Zhao et al., 2024; Liu and Bu, 2024; Chen et al., 2025a; Cui et al., 2026) has emerged as a crucial technique for detecting and attributing machine-generated text. Among various approaches, *distortion-free watermarking* (Aaronson and Kirchner, 2022; Christ et al., 2024; Hu et al., 2024; Wu et al., 2026) has garnered particular attention because it mathematically guarantees that the watermarked text distribution matches the original model distribution in expectation, thereby better preserving generation quality.

Recent advancements have shifted focus from single-layer schemes to *watermark ensembles* (Dathathri et al., 2024; Feng et al., 2025; Wu et al., 2026). By applying watermarking functions recursively across multiple layers, ensemble methods aim to aggregate statistical signals to enhance detectability and robustness. The prevailing design philosophy in these methods typically prioritizes maximizing the strength of the watermark at every individual layer for detectability.

In this work, however, we identify a critical and previously underexplored limitation in this "stronger-is-better" approach. We demonstrate that while strong watermarks improve detectability within the current layer, they inevitably reduce the entropy of the token distribution for subsequent layers. Since the detectability of probabilistic watermarks fundamentally relies on the uncertainty (entropy) of the underlying distribution, this entropy reduction paradoxically weakens the effectiveness of watermarking in later layers. Consequently, we identify an inherent *trade-off* between *single-layer watermark strength* and the *preservation of entropy* required for effective multi-layer ensembling, as illustrated in Fig. 1.

To resolve this dilemma, we propose a counter-intuitive yet theoretically grounded strategy: the use of **weaker single-layer watermarks** to improve overall ensemble performance. By interpolating between the original model distribution and the watermarked distribution using a mixing coefficient λ, we construct a general framework for weaker distortion-free watermarking. We theoretically prove that this approach mitigates the rapid decay of entropy and the expected green-list ratio across layers. By preserving the entropy of the distribution, weaker watermarks allow the ensemble

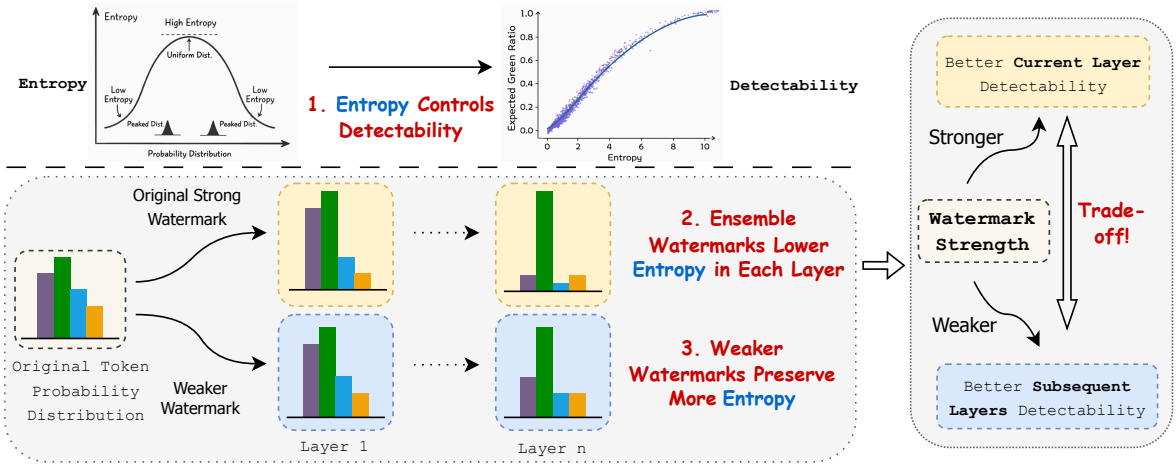


Figure 1: The relationship between entropy, watermark strength, and detectability in distortion-free watermark ensembles. Watermark detectability is closely tied to entropy. Stronger watermarks improve detectability within the current layer. However, they significantly reduce the entropy of the token distribution. In contrast, weaker watermarks preserve more entropy, thereby enhancing detectability in subsequent layers. We propose that there exists an inherent trade-off between the detectability across layers, which can be effectively controlled by adjusting the strength of the watermark.

to accumulate statistical evidence more efficiently over longer sequences, ultimately yielding superior detectability.

We summarize our contributions as follows:

- We demonstrate that for existing watermark methods, detectability is bounded by the entropy of the token distribution.
- We theoretically prove that during watermark ensemble generation, watermarks cause a monotonic decrease in both entropy and detectability across layers.
- We propose a novel framework to achieve better ensemble detectability by preserving more entropy per layer through *weakening* the single-layer watermarking signal. We validate this approach empirically, showing it outperforms strong baselines in both detectability and robustness.

2 Related Work

2.1 Distortion-Free Watermark

To preserve the original output distribution of the language model, several studies have explored distortion-free watermarking techniques. Aaronson and Kirchner (2022) propose an early distortion-free method that applies the Gumbel-max trick to adjust token sampling, using prefix n-grams as watermark keys. Christ et al. (2024) employ inverse sampling to modify token distributions in a binary language model, with watermark keys determined by token positions. ITS-edit and EXP-edit (Kuditipudi et al., 2024) apply inverse sampling and Gumbel-max, respectively, together with a fixed list of watermark keys. Hu et al. (2024) mix inverse sampling with a γ-reweighting strategy. DiPmark (Wu et al., 2024) improves the γ-reweight approach and proposes a robust watermark detector. STA-1 (Mao et al., 2024) focuses on improving text quality in low-entropy generation settings. Dathathri et al. (2024) introduce SynthID, which supports scalable distortion-free watermarking across multiple generations. Finally, Chen et al. (2025c) propose MCmark, which utilizes distribution channels to improve the detectability while remaining distortion-free.

2.2 Watermark Ensemble

Recent research in watermarking for large language models has focused not only on individual watermarking schemes but also on combining multiple watermarking mechanisms to improve detectability and robustness. SynthID (Dathathri et al., 2024) uses multiple layers of random functions to strengthen watermark signals and outperforms simpler single-layer methods in detection performance. We further clarify our contributions compared to SynthID in App. E. BiMark (Feng et al., 2025) achieves robust multi-bit watermarking by ensembling layers of bit-flip unbiased reweighting function based on random bipartitions of the vocabulary. **?** introduces an ensemble framework for distortion-free watermarking, which reduces the amount of text required for detection by increas-

ing the signal-to-noise ratio measured by statistical detectors.

3 Distortion-Free Watermark Ensemble

3.1 Generation

Notation. Let V denote the vocabulary with cardinality $N = |V|$. Given a prompt, a language model M generates tokens autoregressively. At time step t, the probability of generating the next token $x_{t+1} \in V$, conditioned on the preceding sequence $\boldsymbol{x}_{1:t}$, is denoted by $P_M(x_{t+1} \mid \boldsymbol{x}_{1:t}) \in \mathcal{P}$, where $\mathcal{P}$ is the space of probability distributions over V.

Watermarking is performed by reweighting the original distribution $P_M(\cdot \mid \boldsymbol{x}_{1:t})$ into a watermarked distribution $F(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k)$, where $F \in \mathcal{F} : \mathcal{P} \times \mathcal{K} \to \mathcal{P}$ is the watermarking function, $k \in \mathcal{K}$ is a private watermark key sampled from a known distribution $P_{\mathcal{K}}$. The key k is assumed to be exactly recoverable during detection. Here, $\mathcal{K}$ denotes the key space.

Distortion-Free Watermark. Following prior work (Hu et al., 2024; Dathathri et al., 2024), a watermarking function F is said to be *distortion-free* if it preserves the original token distribution in expectation over the key space. Formally, for any $P_M \in \mathcal{P}$ and any token $x_{t+1} \in V$,

$$\begin{aligned} &\mathbb{E}_{k \sim P_{\mathcal{K}}} \left[F(P_M(x_{t+1} \mid \boldsymbol{x}_{1:t}),\, k) \right] \\ =&P_M(x_{t+1} \mid \boldsymbol{x}_{1:t}). \end{aligned} \tag{1}$$

Watermark Ensemble. Following Dathathri et al. (2024); Wu et al. (2026), the detectability of a watermark can be further enhanced by ensembling, which recursively applies the watermarking function F n times using independent watermark keys. Specifically, given n i.i.d. keys $\boldsymbol{k}_{1:n} \sim P_{\mathcal{K}}^n$, the n-fold ensemble transform $\mathrm{ENS} : \mathbb{N} \times \mathcal{P} \times \mathcal{F} \times \mathcal{K}^n \to \mathcal{P}$ is defined recursively as

$$\begin{aligned} &\mathrm{ENS}(n, F, P(\cdot), \boldsymbol{k}_{1:n}) \\ =&\begin{cases} F(\mathrm{ENS}(n-1, F, P(\cdot), \boldsymbol{k}_{1:n-1}), k_n), & n > 1, \\ F(P(\cdot), k_1), & n = 1. \end{cases} \end{aligned} \tag{2}$$

Following the terminology of Dathathri et al. (2024), each application of F corresponds to one *watermark layer*.

If the watermarking function F is distortion-free and the keys $\boldsymbol{k}_{1:n}$ are sampled independently from $P_{\mathcal{K}}$, then the resulting n-layer watermark ensemble remains distortion-free (Dathathri et al., 2024; Wu et al., 2026). That is, for any $P_M(\cdot \mid \boldsymbol{x}_{1:t}) \in \mathcal{P}$,

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{k}_{1:n} \sim P_{\mathcal{K}}^n} \left[\mathrm{ENS}(n, F, P_M(\cdot \mid \boldsymbol{x}_{1:t}), \boldsymbol{k}_{1:n}) \right] \\ =&P_M(\cdot \mid \boldsymbol{x}_{1:t}). \end{aligned} \tag{3}$$

3.2 Detection

Single-Layer Detection. Most existing watermark detection methods (Kirchenbauer et al., 2023; Dathathri et al., 2024) adopt or resemble the red-green list framework introduced by Kirchenbauer et al. (2023). Given a private watermark key k, the vocabulary V is randomly partitioned into a green list V_g^k and a red list V_r^k, such that $|V_g^k| = \gamma |V|$, where $\gamma \in (0, 1)$ is a fixed hyperparameter. During generation, the language model's next-token distribution is modified to bias sampling toward green-list tokens and away from red-list tokens, and have:

$$\begin{aligned} &\sum_{x_{t+1} \in V_g^k} F(P(x_{t+1} | \boldsymbol{x}_{1:t}), k) \\ \geq &\sum_{x_{t+1} \in V_g^k} P(x_{t+1} | \boldsymbol{x}_{1:t}) \end{aligned} \tag{4}$$

Detection is formulated as a hypothesis testing problem, with the null hypothesis H_0 that the sequence was generated without watermarking, and the alternative hypothesis H_1 that the sequence was generated with watermarking. Given a generated sequence $\boldsymbol{x}_{1:T}$ and the corresponding recovered keys $\boldsymbol{k}^{1:T}$, the detector computes the *green ratio*

$$G = \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}\left(x_i \in V_g^{k^i} \right), \tag{5}$$

i.e., the empirical fraction of tokens belonging to the green list. Under the assumption of independent token sampling, the corresponding z-score is computed as

$$z = \frac{(G - \gamma) T}{\sqrt{T \gamma (1 - \gamma)}}. \tag{6}$$

We refer readers to Kirchenbauer et al. (2023) for a detailed derivation and analysis.

Multi-Layer Detection. For an n-layer watermark ensemble, the single-layer detection framework can be naturally extended under the assumption that watermark keys across layers are independently and identically distributed. Let G_i denote the green ratio associated with the i-th layer:

$$G_i = \frac{1}{T}\sum_{j=1}^{T} \mathbb{1}\left(x_j \in V_g^{k_i^j}\right), \quad i = 1, \ldots, n. \tag{7}$$

Aggregating across all layers, the corresponding multi-layer z-score is given by

$$z = \frac{\left(\sum_{i=1}^{n} G_i - \gamma n\right) T}{\sqrt{nT\gamma(1-\gamma)}}. \tag{8}$$

Finally, for a single-layer watermark F, we characterize the expected probability that the next generated token belongs to the green list under the original model distribution $P_M(\cdot \mid \boldsymbol{x}_{1:t})$. We define the *expected green ratio* function $g_F : \mathcal{P} \to [0, 1]$ as

$$\begin{aligned} &g_F(P_M(\cdot \mid \boldsymbol{x}_{1:t})) \\ =&\mathbb{E}_{k\sim P_{\mathcal{K}}} \sum_{x_{t+1}\in V_g^k} F(P_M(x_{t+1} \mid \boldsymbol{x}_{1:t}), k)\,, \end{aligned} \tag{9}$$

where the expectation is taken over the watermark key distribution $P_{\mathcal{K}}$.

4 Rethinking Distortion-Free Ensemble Watermarks from an Entropy Perspective

4.1 Larger Entropy Leads to Better Detection Performance

It is well established that the detectability of probabilistic watermarks depends critically on the uncertainty of the original token distribution prior to watermarking (Hu et al., 2024; Dathathri et al., 2024; Giboulot and Furon, 2024). Intuitively, when the model output distribution exhibits higher entropy, watermark-induced perturbations are less likely to be masked by deterministic biases of the base model, thereby yielding more reliable statistical signals for detection.

We formalize this intuition by analyzing the expected green-list ratio for each watermark method. Under the SynthID watermarking scheme (Dathathri et al., 2024), we show that the expected green ratio under SynthID can be expressed as

$$\begin{aligned} &g_{\text{SynthID}}(P_M(\cdot \mid \boldsymbol{x}_{1:t})) \\ =&\frac{3}{4} - \frac{1}{4}\sum_{x_{t+1}\in V} \left(P_M(x_{t+1} \mid \boldsymbol{x}_{1:t})\right)^2 \\ \leq&\frac{3}{4} - \frac{1}{4}\exp\big(-H(P_M(\cdot \mid \boldsymbol{x}_{1:t}))\big), \end{aligned} \tag{10}$$

where $H(P_M(\cdot \mid \boldsymbol{x}_{1:t})) = -\sum_{x_{t+1}\in V} P_M(x_{t+1} \mid \boldsymbol{x}_{1:t}) \log P_M(x_{t+1} \mid \boldsymbol{x}_{1:t})$ denotes the Shannon entropy of the model's output distribution. A detailed derivation is provided in Appendix A.

For other distortion-free ensemble watermarking schemes, deriving an analogous closed-form expression is generally intractable due to more complex reweighting mechanisms. Nevertheless, in Sec. 6.2, we empirically demonstrate a consistent and strong positive correlation between the entropy of the base distribution and the expected green ratio across a variety of watermark designs, supporting the generality of this entropy-based perspective.

4.2 Watermarks Reduce Entropy and Expected Green Ratio in Subsequent Layers

We theoretically identify an important phenomenon: *distortion-free watermarking inevitably reduces both the entropy of the token distribution and the expected green ratio in subsequent layers*. This effect compounds across layers, progressively weakening watermark detectability.

Theorem 4.1 (Entropy Decrease under Distortion-Free Watermarking)**.** *Let F denote a distortion-free watermarking operator with private key $k \sim P_{\mathcal{K}}$. Then, in expectation over the watermark key, the Shannon entropy of the token distribution after watermarking does not increase:*

$$\begin{aligned} &\mathbb{E}_{k\sim P_{\mathcal{K}}}\left[H(F(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k))\right] \\ \leq&H(P_M(\cdot \mid \boldsymbol{x}_{1:t})). \end{aligned} \tag{11}$$

Proof. Since Shannon entropy is a concave functional over the probability simplex, Jensen's inequality yields

$$\begin{aligned} &H(P_M(\cdot \mid \boldsymbol{x}_{1:t})) \\ =&H\left(\mathbb{E}_{k\sim P_{\mathcal{K}}}\left[F(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k)\right]\right) \\ \geq&\mathbb{E}_{k\sim P_{\mathcal{K}}}\left[H(F(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k))\right], \end{aligned} \tag{12}$$

which completes the proof. □

Theorem 4.2 (Expected Green Ratio Decrease)**.** *Under expectation over the watermark key, the expected green ratio does not increase after applying a distortion-free watermark:*

$$\begin{aligned} &\mathbb{E}_{k\sim P_{\mathcal{K}}}\left[g(F(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k))\right] \\ \leq&g(P_M(\cdot \mid \boldsymbol{x}_{1:t})). \end{aligned} \tag{13}$$

Proof sketch. The argument parallels Theorem 4.1. If the expected green ratio functional $g(\cdot)$ is concave with respect to the token distribution, then Jensen's inequality directly implies the result.

For SynthID, Eq. 10 admits a closed-form expression which is concave. For other representative

distortion-free watermarking schemes, we also theoretically verify concavity of their expected green-ratio functions in Appendix A. □

Implication. Taken together, Theorems 4.1 and 4.2 imply that watermark ensemble generation inevitably induces a monotonic reduction in both entropy and expected green ratio across successive layers. As a consequence, the statistical signal exploited by downstream detectors progressively diminishes, fundamentally limiting watermark detectability.

4.3 A General Weaker Distortion-Free Watermark Framework

A direct approach to mitigating the degradation of watermark detectability across layers is to reduce the entropy loss induced at each step, thereby maintaining a larger expected green ratio in subsequent layers. This motivates the design of a *weaker* distortion-free watermark that perturbs the base token distribution more mildly.

Concretely, we construct a family of weaker watermarking functions by interpolating between the original model distribution and a distortion-free watermarked distribution. Let F denote a distortion-free watermarking function. For a mixing coefficient $\lambda \in [0, 1]$, we define

$$\begin{aligned} &F_\lambda(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k) \\ &:= \lambda\, F(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k) + (1-\lambda)\, P_M(\cdot \mid \boldsymbol{x}_{1:t}). \end{aligned} \tag{14}$$

The parameter λ controls the watermark strength. $\lambda = 1$ recovers the original watermark F, while smaller values of λ yield distributions closer to the base model and thus induce smaller perturbations. We show that this interpolation preserves distortion-freeness and in expectation improves entropy relative to directly applying F, which in turn helps sustain detectability in later layers.

Theorem 4.3 (Distortion-Freeness)**.** *If F is distortion-free, then for any $\lambda \in [0, 1]$, F_λ is also distortion-free. In particular, for any token x_{t+1},*

$$\begin{aligned} &\mathbb{E}_{k\sim P_{\mathcal{K}}}\Big[F_\lambda(P_M(x_{t+1} \mid \boldsymbol{x}_{1:t}), k)\Big] \\ =&P_M(x_{t+1} \mid \boldsymbol{x}_{1:t}). \end{aligned} \tag{15}$$

By linearity of expectation, the proof is straightforward, and we leave it in Appendix B.

Theorem 4.4 (Entropy Preservation)**.** *Assume F is distortion-free. Then for any $\lambda \in [0, 1]$,*

$$\begin{aligned} &\mathbb{E}_{k\sim P_{\mathcal{K}}}\Big[H\big(F_\lambda(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k)\big)\Big] \\ \geq\ &\mathbb{E}_{k\sim P_{\mathcal{K}}}\Big[H\big(F(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k)\big)\Big]. \end{aligned} \tag{16}$$

Proof Sketch. For each fixed key k, $F_\lambda(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k)$ is a convex combination of $F(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k)$ and $P_M(\cdot \mid \boldsymbol{x}_{1:t})$. Since Shannon entropy is concave on the probability simplex,

$$H\big(F_\lambda(\cdot)\big) \ \geq\ \lambda\, H\big(F(\cdot)\big) + (1-\lambda)\, H\big(P_M(\cdot)\big).$$

Taking expectation over k and using distortion-freeness to relate P_M to the key-average behavior yields the stated inequality. Full details are deferred to Appendix B. □

The proposed interpolation in (14) offers a principled framework for modulating watermark strength without violating the distortion-free constraint. In this context, we observe a critical trade-off between immediate detectability and long-term sustainability: maximizing the watermark signal at the current step often collapses the entropy necessary for embedding signals in future steps. Our proposed weaker watermark, F_λ, directly addresses this tension. By reducing entropy degradation at the current step, F_λ preserves a higher expected green ratio for subsequent layers, effectively counteracting the phenomenon of detectability decay.

5 Experiments

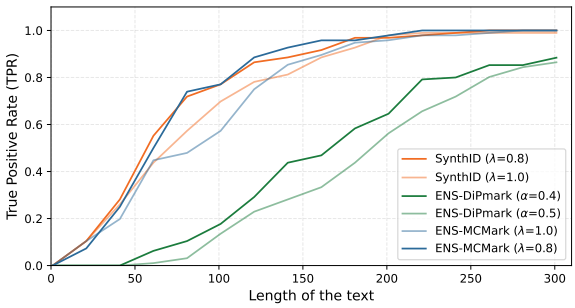


Figure 2: Watermark detectability as a function of text length on the MMW Story dataset using Llama-3.2-3B-Instruct. The threshold for false positive rate is set to 0.01%. Weaker ensemble watermarks yield consistently higher detection performance.

5.1 Settings and Hyperparameters

For distortion-free watermarking, we leverage SynthID (Dathathri et al., 2024), ENS-DiPmark (Wu et al., 2024, 2026), and ENS-MCMark (Chen et al.,

Table 1: Watermark detection performance with Llama-3.2-3B-Instruct and Mistral-7B-Instruct-v0.3 on the C4 dataset under different token budgets. We report TPR at fixed FPR thresholds (0.1%, 0.01%, 0.001%) and the median detection p-value across samples. Lower p-values and higher TPR indicate stronger detectability.

		150 tokens				250 tokens			
		TPR@FPR			Median p-value↓	TPR@FPR			Median p-value↓
		0.1%↑	0.01%↑	0.001%↑		0.1%↑	0.01%↑	0.001%↑	
Llama3-3B	ENS-DiPmark (α=0.5)	52.8%	42.9%	35.2%	7.34e-4	74.0%	64.1%	54.9%	2.04e-6
	ENS-DiPmark (α=0.4)	**55.5%**	**45.2%**	**38.2%**	**4.20e-4**	**74.7%**	**67.2%**	**58.5%**	**5.36e-7**
	SynthID (λ=1)	75.1%	68.3%	61.9%	1.33e-7	93.8%	88.4%	83.5%	1.91e-12
	SynthID (λ=0.8)	**79.7%**	**74.5%**	**67.1%**	**1.57e-8**	**94.5%**	**90.5%**	**86.1%**	**1.69e-13**
	ENS-MCMark (λ=1)	77.2%	71.4%	64.1%	**7.80e-9**	95.1%	91.8%	87.3%	8.75e-15
	ENS-MCMark (λ=0.8)	**78.5%**	**71.6%**	**65.2%**	**7.80e-9**	**96.7%**	**93.7%**	**88.5%**	**3.91e-15**
Mistral-7B	ENS-DiPmark (α=0.5)	22.6%	11.7%	6.3%	2.60e-2	46.8%	27.8%	19.1%	2.13e-3
	ENS-DiPmark (α=0.4)	**25.4%**	**13.7%**	**6.4%**	**1.37e-2**	**48.3%**	**33.6%**	**20.7%**	**1.42e-3**
	SynthID (λ=1)	73.1%	58.8%	45.8%	2.06e-5	91.9%	83.2%	75.0%	3.23e-8
	SynthID (λ=0.8)	**78.8%**	**64.3%**	**49.4%**	**1.06e-5**	**93.6%**	**87.4%**	**78.0%**	**2.19e-8**
	ENS-MCMark (λ=1)	78.5%	67.1%	52.4%	9.11e-6	93.3%	87.5%	80.5%	7.32e-9
	ENS-MCMark (λ=0.8)	**82.2%**	**72.2%**	**57.6%**	**2.50e-6**	**93.6%**	**88.8%**	**82.1%**	**2.03e-9**

2025c; Wu et al., 2026). We use the default hyperparameter settings for each method: for SynthID, we set $m = 30$; for ENS-DiPmark, we set $n = 5$; and for ENS-MCMark, we set $n = 5$ and $l = 20$, resulting in 30 layers for SynthID and 5 layers for both ENS-DiPmark and ENS-MCMark. To further explore the effect of the number of layers, we conduct an ablation study in Appendix D.2.

Following prior works, we report the True Positive Rate (TPR) at specific controlled False Positive Rates (FPR), using values of 0.1%, 0.01%, and 0.001% in our experiments. Additionally, we present the median p-values for the dataset.

Regarding watermark strength, ENS-DiPmark includes a hyperparameter α that controls the watermark strength. We use $\alpha = 0.5$ and $\alpha = 0.4$ for our experiments, as $\alpha = 0.5$ yields the best performance for single-layer DiPmark. For SynthID and ENS-MCMark, since no explicit hyperparameter is available to control watermark strength, we employ our proposed weaker watermark framework, as defined in Eq. (14). In our main experiments, we test $\lambda = 1$ and $\lambda = 0.8$. Details of models and datasets used are shown in Appendix C.

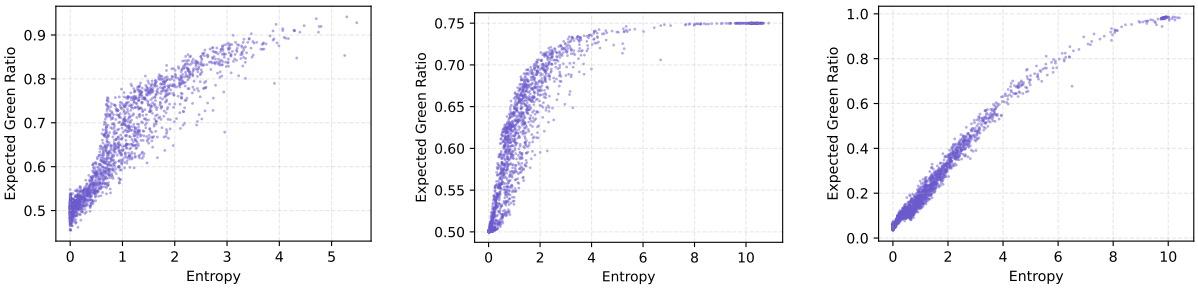
5.2 Experimental Results

We report watermark detection performance across two backbone models and two generation lengths (150 and 250 tokens) in Tab. 1. For distortion-free ensemble methods, weakening the watermark strength (reducing α in ENS-DiPmark or λ in ENS-MCMark and SynthID) consistently improves detection performance, indicating that milder perturbations preserve more entropy and yield stronger downstream statistical evidence. Fig. 2 illustrates how watermark detectability evolves as a function of text length. We observe that weaker watermark configurations consistently dominate their stronger counterparts, highlighting the benefit of preserving token entropy. Additional results on more datasets are provided in App. D.1 in Fig. 7. These trends are consistent across models, datasets and watermark schemes, demonstrating the robustness and generality of the proposed weaker ensemble watermarking framework.

6 Analysis

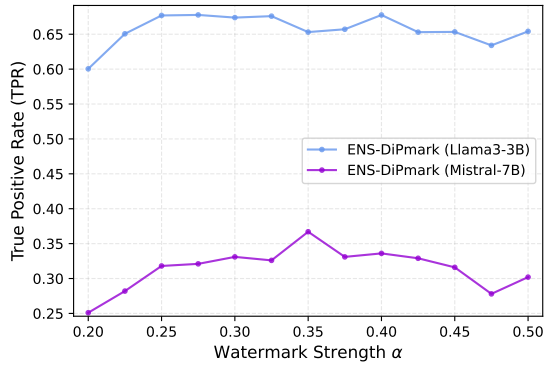
6.1 Effect of Watermark Strength

We study the effect of watermark strength on detection performance in Fig. 4. Decreasing the watermark strength generally improves the true positive rate, with peak performance attained around λ=0.8 and α=0.4. This behavior indicates that excessively strong perturbations can be detrimental, as they induce excessive entropy reduction that undermines downstream detectability. Importantly, the strength parameters α and λ explicitly govern the trade-off between detectability at the current layer and en-

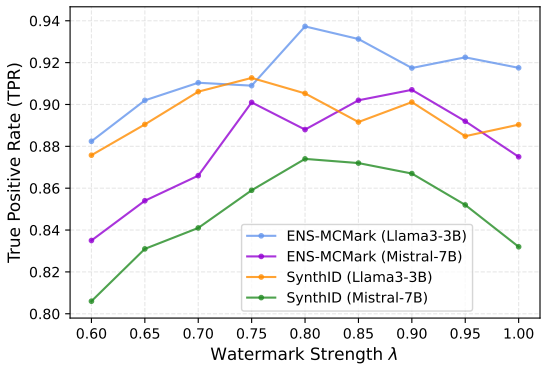


(a) ENS-DiPmark (b) SynthID (c) ENS-MCMark

Figure 3: Correlation between token-distribution entropy and expected green ratio. Results are computed on the MMW Story dataset with Llama3.2-3B-Instruct using 2000 randomly sampled tokens, showing a strong positive association between entropy and expected green ratio.



(a) ENS-DiPmark



(b) SynthID and ENS-MCMark

Figure 4: Ablation of watermark strength on detection performance on the C4 dataset with false positive rate set to 0.01% and token length set to 250. True positive rate (TPR) is reported for varying strength parameters α (ENS-DiPmark) and λ (SynthID, ENS-MCMark) on Llama3-3B and Mistral-7B. Moderate weakening consistently yields the best detectability.

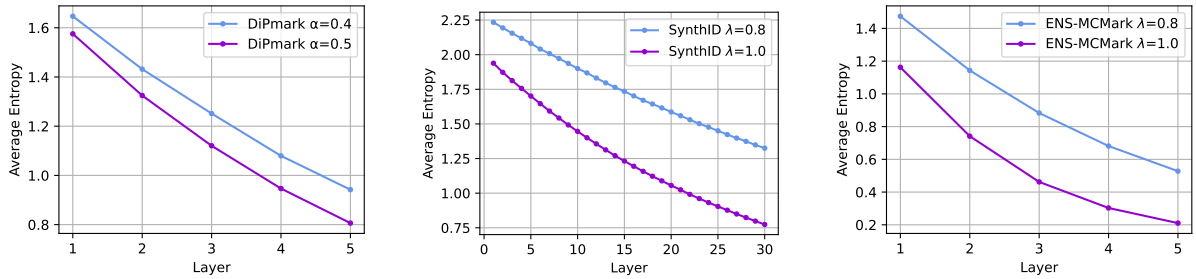
tropy preservation, which in turn affects detectability in subsequent layers. These results provide empirical evidence supporting our central claim that weaker distortion-free watermarking achieves a more favorable balance between immediate detection strength and overall detectability.

6.2 Correlation between Entropy and Expected Green Ratio

We visualize the empirical relationship between the entropy of the pre-watermark token distribution and the resulting expected green ratio for three watermarking schemes in Fig. 3. Across ENS-DiPmark, SynthID, and ENS-MCMark, we observe a strong positive association: higher-entropy distributions yield systematically larger expected green ratios, whereas low-entropy distributions compress the green ratio toward its baseline level, reducing the statistical margin available for detection. The results provide empirical evidence for our entropy-based perspective: entropy serves as a key determinant of watermark detectability through its control of the expected green ratio.

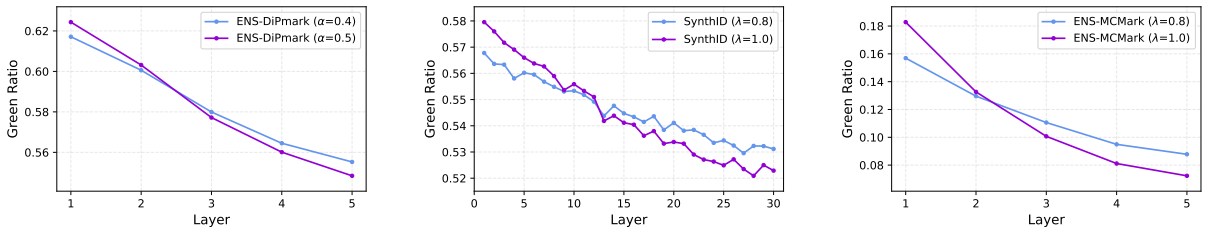
6.3 Entropy and Green Ratio in Each Layer

We empirically analyze how watermark strength influences the evolution of entropy and green ratio across layers. Fig. 5 shows that, for all methods, the token distribution entropy decreases monotonically for each layer. Importantly, weaker watermark settings (smaller α and λ) consistently preserve higher entropy at each layer, slowing entropy collapse for all the layers. Fig. 6 further demonstrates that this entropy preservation directly translates into higher expected green ratios in later layers, while stronger settings suffer from accelerated degradation. Together, these results provide empirical support for our theoretical analysis in Sec. 4.2, confirming that watermark strength controls a fundamental trade-off between immediate detectability and long-horizon statistical capacity, and that weaker distortion-free watermarking better preserves entropy and green ratio, thereby enabling stronger detectability in subsequent layers.



(a) ENS-DiPmark (b) SynthID (c) ENS-MCMark

Figure 5: Average token entropy per layer under different watermark strengths on the C4 dataset using Llama-3.2-3B-Instruct with sequence length fixed to 150. Weaker watermark settings consistently preserve higher entropy across layers.



(a) ENS-DiPmark (b) SynthID (c) ENS-MCMark

Figure 6: Actual green ratio G_i per layer under different watermark strengths on the C4 dataset using Llama-3.2-3B-Instruct with sequence length fixed to 150. Higher entropy preservation from weaker watermarking leads to larger green ratios in later layers.

Table 2: Robustness of watermark detection under four attacks with different watermark strength. We report TPR (at FPR=0.1%) and median p-value on the C4 dataset using Llama3.2-3B-Instruct with sequence length 250.

	Random Token Replacement		GPT Back Translation		GPT Rephrase		DIPPER	
	TPR↑	Median p-value↓	TPR↑	Median p-value↓	TPR↑	Median p-value↓	TPR↑	Median p-value↓
ENS-DiPmark (α=0.5)	54.2%	4.92e-4	18.6%	6.79e-2	1.9%	5.27e-1	1.6%	6.30e-1
ENS-DiPmark (α=0.4)	**55.9%**	**3.14e-4**	**22.4%**	**3.92e-2**	**2.5%**	**4.61e-1**	**2.0%**	**5.95e-1**
SynthID (λ=1)	81.1%	9.48e-8	51.8%	6.90e-4	12.7%	6.82e-2	8.7%	9.59e-2
SynthID (λ=0.8)	**84.8%**	**3.44e-8**	**54.2%**	**4.99e-4**	**13.0%**	**6.52e-2**	**10.3%**	**9.21e-2**
ENS-MCMark (λ=1)	85.9%	73.2e-9	55.3%	2.97e-4	13.8%	6.27e-2	12.3%	9.92e-2
ENS-MCMark (λ=0.8)	**87.5%**	**3.88e-9**	**58.9%**	**1.24e-4**	**16.4%**	**3.78e-2**	**12.7%**	**7.93e-2**

6.4 Robustness Experiments

We evaluate the robustness of our proposed weaker watermark framework in Tab. 2 under four attacks, including random token replacement, GPT-based back translation, GPT rephrasing (Sadasivan et al., 2023), and DIPPER (Krishna et al., 2023) paraphrasing attack, following the settings of Wu et al. (2026). The results further demonstrate that our weaker distortion-free watermark framework does not sacrifice robustness. Instead, it consistently improves detectability under attacks. Across all four perturbations, the weaker configurations achieve higher TPR and smaller median p-values, indicating stronger statistical evidence.

7 Conclusion

In this work, we identify a critical limitation in existing watermark ensembles where maximizing single-layer strength paradoxically degrades long-term detectability by reducing the entropy of the token distribution. To address this trade-off, we propose a general framework that utilizes

weaker distortion-free watermarks to preserve entropy across layers, thereby maintaining a higher expected green ratio for subsequent layers. Our theoretical proofs and empirical evaluations confirm that this entropy-preserving strategy significantly outperforms strong baselines, achieving superior detectability and robustness in multi-layer settings.

Limitations

While our work demonstrates the effectiveness of weaker distortion-free watermarking ensembles, there are a few limitations to consider. First, our experimental evaluation was only conducted on open-source models with sizes ranging from 3B to 7B parameters.

Additionally, we did not explore adaptive strategies that dynamically tune the watermark strength based on the real-time entropy of the token distribution, which could potentially further optimize the trade-off between detectability and signal preservation.

Acknowledgments

This work was partially supported by NSF IIS 2347592, 2348169, DBI 2405416, CCF 2348306, CNS 2347617, RISE 2536663.

References

Scott Aaronson and Hendrik Kirchner. 2022. Watermarking of large language models.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Ruibo Chen, Yihan Wu, Yanshuo Chen, Chenxi Liu, Junfeng Guo, and Heng Huang. 2025a. A watermark for order-agnostic language models. In *International Conference on Learning Representations*, volume 2025, pages 4372–4388.

Ruibo Chen, Yihan Wu, Junfeng Guo, and Heng Huang. 2025b. De-mark: Watermark removal in large language models. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net.

Ruibo Chen, Yihan Wu, Junfeng Guo, and Heng Huang. 2025c. Improved unbiased watermark for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 20587–20601. Association for Computational Linguistics.

Miranda Christ, Sam Gunn, and Or Zamir. 2024. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory, June 30 - July 3, 2023, Edmonton, Canada*, Proceedings of Machine Learning Research, pages 1125–1139. PMLR.

Xuehao Cui, Ruibo Chen, Yihan Wu, and Heng Huang. 2026. Mc2mark: Distortion-free multi-bit watermarking for long messages. *arXiv preprint arXiv:2602.14030*.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, and 1 others. 2024. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xiaoyan Feng, He Zhang, Yanjun Zhang, Leo Yu Zhang, and Shirui Pan. 2025. Bimark: Unbiased multilayer watermarking for large language models. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Eva Giboulot and Teddy Furon. 2024. Watermax: breaking the llm watermark detectability-robustness-quality trade-off. *Advances in Neural Information Processing Systems*, 37:18848–18881.

Josh A Goldstein. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. 2024. Unbiased watermark for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36:27469–27500.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2024. Robust distortion-free watermarks for language models. *Trans. Mach. Learn. Res.*, 2024.

Yepeng Liu and Yuheng Bu. 2024. Adaptive text watermark for large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 30718–30737. PMLR / OpenReview.net.

Minjia Mao, Dongjun Wei, Zeyu Chen, Xiao Fang, and Michael Chau. 2024. A watermark for low-entropy and unbiased generation in large language models. *arXiv preprint arXiv:2405.14604*.

Julien Piet, Chawin Sitawarin, Vivian Fang, Norman Mu, and David Wagner. 2023. Mark my words: Analyzing and evaluating language model watermarks. *arXiv preprint arXiv:2312.00273*.

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

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Shangqing Tu, Yuliang Sun, Yushi Bai, Jifan Yu, Lei Hou, and Juanzi Li. 2024. Waterbench: Towards holistic evaluation of watermarks for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1517–1542. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Yihan Wu, Ruibo Chen, Georgios Milis, and Heng Huang. 2026. An ensemble framework for unbiased language model watermarking. In *International Conference on Learning Representations*, volume 2026, pages 17899–17917.

Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. 2024. A resilient and accessible distribution-preserving watermark for large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 53443–53470. PMLR / OpenReview.net.

Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu-Xiang Wang. 2024. Provable robust watermarking for ai-generated text. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

A Expected Green Ratio for Single-Layer Distortion-Free Watermarks

In this section, we provide the theoretical proofs for the concavity of the expected green ratio function $g(\cdot)$ and its relationship with entropy for the discussed watermarking schemes. For brevity, we denote $P_M(\cdot|x_{1:t})$ as $P(\cdot)$ when the context is unambiguous.

A.1 SynthID

Expected Green Ratio. As established in prior work (Dathathri et al., 2024), the expected green ratio for the SynthID scheme is given by:

$$g_{\text{SynthID}}(P) = \frac{3}{4} - \frac{1}{4}\sum_{x\in V}(P(x))^2.$$

Proof of Concavity. We aim to show that $g_{\text{SynthID}}(P)$ is a concave function over the probability simplex $\mathcal{P}$. Let $f(P) = \sum_{x\in V}(P(x))^2$. This represents the collision probability (or the squared L_2 norm) of the distribution. Consider the function $h(y) = y^2$, which is strictly convex for $y \in \mathbb{R}$. Since $f(P)$ is a sum of convex functions applied to the components of P, $f(P)$ is convex with respect to P.The expected green ratio is an affine transformation of $f(P)$ with a negative coefficient:

$$g_{\text{SynthID}}(P) = \frac{3}{4} - \frac{1}{4}f(P).$$

Since multiplying a convex function by a negative constant reverses the inequality, $-\frac{1}{4}f(P)$ is concave. Adding a constant ($\frac{3}{4}$) preserves concavity. Therefore, $g_{\text{SynthID}}(P)$ is concave.

Relation with Entropy. (Proof of Eq. 9) We provide the derivation for the bound:

$$g_{\text{SynthID}}(P) \le \frac{3}{4} - \frac{1}{4}\exp(-H(P)).$$

Recall the definition of Shannon entropy $H(P) = -\sum_{x\in V} P(x)\ln P(x)$.We utilize the relationship between the collision probability and

entropy. By Jensen's inequality applied to the concave function $\ln(\cdot)$:

$$\begin{aligned}&\ln\left(\sum_{x\in V} P(x)^2\right)\\ =&\ln\left(\mathbb{E}_{x\sim P}[P(x)]\right)\\ \geq&\mathbb{E}_{x\sim P}[\ln P(x)].\end{aligned} \tag{17}$$

Substituting the definition of entropy:

$$\mathbb{E}_{x\sim P}[\ln P(x)] = \sum_{x\in V} P(x)\ln P(x) = -H(P).$$

Thus, we have:

$$\ln\left(\sum_{x\in V} P(x)^2\right) \geq -H(P).$$

Exponentiating both sides (since e^x is monotonically increasing):

$$\sum_{x\in V} P(x)^2 \geq \exp(-H(P)).$$

Now, substituting this inequality back into the expected green ratio equation:

$$g_{\text{SynthID}}(P) = \frac{3}{4} - \frac{1}{4}\sum_{x\in V}(P(x))^2.$$

Multiplying the inequality by $-\frac{1}{4}$ reverses the sign:

$$-\frac{1}{4}\sum_{x\in V}(P(x))^2 \leq -\frac{1}{4}\exp(-H(P)).$$

Finally, adding $\frac{3}{4}$ to both sides yields:

$$g_{\text{SynthID}}(P) \leq \frac{3}{4} - \frac{1}{4}\exp(-H(P)).$$

This confirms that the expected green ratio is upper-bounded by a function of the entropy.

A.2 DiPmark

Expected Green Ratio. For DiPmark (Wu et al., 2024), the expected green ratio is:

$$g_{\text{DiPmark}}(P) = \mathbb{E}_g\left[\min\left(1, \sum_{i=1}^{|V|} P(x_i)g_i + \alpha\right)\right],$$

where $g_i \in \{0,1\}$ determines the partition of the vocabulary, $\sum_i g_i = \frac{|V|}{2}$ and α is a shift parameter.

Proof of Concavity. We show that $g_{\text{DiPmark}}(P)$ is concave with respect to P.

Let $L(P,g) = \sum_{i=1}^{|V|} P(x_i)g_i + \alpha$. Note that for a fixed key/partition g, $L(P,g)$ is a linear function of P. Define the function $h(y) = \min(1,y)$. The function $h(y)$ is the minimum of two concave functions. Thus, $h(y)$ is concave. Since the composition of a concave function with a linear function is concave, the term inside the expectation:

$$\begin{aligned}&\phi(P,g) = h(L(P,g))\\ =&\min\left(1, \sum_{i=1}^{|V|} P(x_i)g_i + \alpha\right),\end{aligned} \tag{18}$$

is concave with respect to P for any fixed g. The expected green ratio $g_{\text{DiPmark}}(P)$ is the expectation of $\phi(P,g)$ over the distribution of watermark keys g. Since expectation is a linear operator with non-negative weights (probabilities), and a non-negative weighted sum of concave functions preserves concavity, $g_{\text{DiPmark}}(P)$ is concave.

A.3 MCMark

Expected Green Ratio. For MCMark (Chen et al., 2025c), with channel number l, the expected green ratio is:

$$g_{\text{MCMark}}(P) = \mathbb{E}_g\left[\min\left(1, l\sum_{i=1}^{|V|} P(x_i)g_i\right)\right],$$

where $g_i \in \{0,1\}$ determines the partition of the vocabulary, $\sum_i g_i = \frac{|V|}{l}$.

Proof of Concavity. The proof follows the same logic as DiPmark.

Let $K(P,g) = l\sum_{i=1}^{|V|} P(x_i)g_i$. This term represents the scaled probability mass falling into a specific channel determined by g. $K(P,g)$ is strictly linear with respect to the distribution P. Consider the function $h(y) = \min(1,y)$, which is a concave function. Therefore, the composition $\psi(P,g) = h(K(P,g)) = \min(1, l\sum P(x_i)g_i)$ is concave with respect to P.

Finally, taking the expectation over g:

$$g_{\text{MCMark}}(P) = \mathbb{E}_g[\psi(P,g)].$$

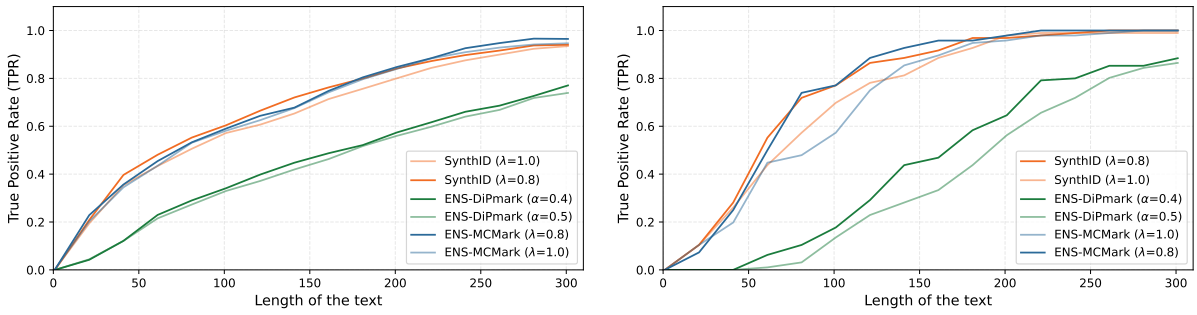
Since the expectation preserves concavity, $g_{\text{MCMark}}(P)$ is a concave function of the token distribution P.

Table 3: Ablation study on the number of ensemble layers for ENS-DiPmark using the Llama3-3B and Mistral-7B models. Results are reported as the true positive rate at a false positive rate of 0.01% with 250 tokens per sample on the C4 dataset.

Model	Watermark	Layer Number					
		1	2	3	4	5	Best
Llama3-3B	ENS-DiPmark (α=0.5)	**40.4%**	**59.0%**	63.7%	63.9%	64.1%	64.1%
	ENS-DiPmark (α=0.4)	38.57%	57.2%	**63.9%**	**66.9%**	**67.2%**	**67.2%**
Mistral-7B	ENS-DiPmark (α=0.5)	**11.1%**	**29.5%**	30.4%	30.7%	27.8%	30.7%
	ENS-DiPmark (α=0.4)	9.0%	28.2%	**33.0%**	**35.5%**	**33.6%**	**35.5%**

Table 4: Ablation study on the number of ensemble layers for SynthID using the Llama3-3B and Mistral-7B models. Results are reported as the true positive rate at a false positive rate of 0.01% with 250 tokens per sample on the C4 dataset.

Model	Watermark	Layer Number						
		5	10	15	20	25	30	Best
Llama3-3B	SynthID (λ=1)	**72.2%**	**85.4%**	**88.0%**	88.2%	88.6%	88.4%	88.6%
	SynthID (λ=0.8)	60.5%	79.8%	85.8%	**89.1%**	**90.4%**	**90.5%**	**90.5%**
Mistral-7B	SynthID (λ=1)	**59.0%**	**81.4%**	**85.2%**	**86.0%**	**85.5%**	83.2%	86.0%
	SynthID (λ=0.8)	36.1%	69.6%	78.7%	82.9%	85.3%	**87.4%**	**87.4%**



(a) C4 (b) MMW Story

Figure 7: Watermark detectability as a function of text length on C4 and Longform QA using Llama-3.2-3B-Instruct. The threshold for false positive rate is set to 0.01%. Weaker ensemble watermarks yield consistently higher detection performance across datasets.

B A General Weaker Distortion-Free Watermark Framework

B.1 Proof for Theorem 4.3

Proof. By definition of F_λ and linearity of expectation, we have

$$\begin{aligned}&\mathbb{E}_{k\sim P_\mathcal{K}}\left[F_\lambda(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k)\right]\\ =&\lambda\mathbb{E}_{k\sim P_\mathcal{K}}\left[F(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k)\right]\\ &+(1-\lambda)P_M(\cdot \mid \boldsymbol{x}_{1:t}).\end{aligned} \tag{19}$$

Since F is distortion-free by assumption, $\mathbb{E}_{k\sim P_\mathcal{K}}[F(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k)] = P_M(\cdot \mid \boldsymbol{x}_{1:t})$, which immediately yields $\mathbb{E}_{k\sim P_\mathcal{K}}[F_\lambda(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k)] = P_M(\cdot \mid \boldsymbol{x}_{1:t})$. □

B.2 Proof for Theorem 4.4

Proof. Fix a key k. By construction, $F_\lambda(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k)$ is a convex combination of the two distributions $F(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k)$ and $P_M(\cdot \mid \boldsymbol{x}_{1:t})$. Since Shannon entropy is a concave function over the probability simplex, we have

$$\begin{aligned}&H(F_\lambda(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k))\\ =&H\big(\lambda F(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k) + (1-\lambda)P_M(\cdot \mid \boldsymbol{x}_{1:t})\big)\\ \geq&\lambda H(F(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k))\\ &+(1-\lambda)H(P_M(\cdot \mid \boldsymbol{x}_{1:t})).\end{aligned} \tag{20}$$

Table 5: Ablation study on the number of ensemble layers for ENS-MCMark using the Llama3-3B and Mistral-7B models. Results are reported as the true positive rate at a false positive rate of 0.01% with 250 tokens per sample on the C4 dataset.

Model	Watermark	Layer Number					
		1	2	3	4	5	Best
Llama3-3B	ENS-MCMark (λ=1)	**86.6%**	**91.1%**	92.0%	92.4%	91.8%	92.4%
	ENS-MCMark (λ=0.8)	77.1%	87.8%	**92.6%**	**93.6%**	**93.7%**	**93.7%**
Mistral-7B	ENS-MCMark (λ=1)	**78.4%**	**86.6%**	**88.1%**	**87.5%**	87.5%	88.1%
	ENS-MCMark (λ=0.8)	57.8%	79.2%	85.0%	86.2%	**88.8%**	**88.8%**

Taking expectation over $k \sim P_{\mathcal{K}}$ yields

$$\begin{aligned} &\mathbb{E}_k\big[H(F_\lambda(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k))\big] \\ \geq &\lambda \mathbb{E}_k\big[H(F(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k))\big] \\ &+ (1-\lambda) H(P_M(\cdot \mid \boldsymbol{x}_{1:t})). \end{aligned} \tag{21}$$

Finally, since F is distortion-free, Jensen's inequality implies $\mathbb{E}_k[H(F(P_M(\cdot \mid \boldsymbol{x}_{1:t}), k))] \leq H(P_M(\cdot \mid \boldsymbol{x}_{1:t}))$. Substituting this bound into the inequality above gives $\mathbb{E}_k[H(F_\lambda(\cdot))] \geq \mathbb{E}_k[H(F(\cdot))]$, which completes the proof. □

C Models and Datasets

In our experiments, we follow prior works (Chen et al., 2025c; Wu et al., 2026) and utilize several datasets, including a subset of the C4 dataset (Raffel et al., 2020), the MMW Story dataset (Piet et al., 2023), and the Longform QA dataset from WaterBench (Tu et al., 2024). For text generation, we employ language models such as Llama-3.2-3B-Instruct (Dubey et al., 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023).

We run our experiments on 4 NVIDIA RTX 6000 Ada GPUs.

D Additional Experiments

D.1 Watermark Detectability under Different Length

Fig. 2 and 7 illustrates how watermark detectability evolves as a function of text length across different datasets. We observe that weaker watermark configurations consistently dominate their stronger counterparts, highlighting the benefit of preserving token entropy. These trends are consistent across models, datasets and watermark schemes, demonstrating the robustness and generality of the proposed weaker ensemble watermarking framework.

D.2 Ablation on the Number of Ensemble Layers

To mitigate the effect of the number of layers, we evaluate the True Positive Rate (TPR) across a varying number of ensemble layers for ENS-DiPmark, SynthID, and ENS-MCMark on both the Llama3-3B and Mistral-7B models.

As shown in Table 3,4 and 5, stronger watermarks consistently exhibit superior detectability in the initial layers. However, the performance of these stronger watermarks quickly plateaus or even degrades as the number of layers increases due to rapid entropy collapse. In contrast, weaker watermarks preserve more entropy, allowing the ensemble to accumulate reliable statistical signals over a larger number of layers. Across all three methods, the weaker configurations ultimately overtake their stronger counterparts in later layers, consistently achieving a higher maximum detectability when an adequate number of ensemble layers is employed.

D.3 Results on Larger Models

To verify that the benefit of weaker ensemble watermarking is not limited to the 3B–7B backbones used in our main experiments, we further evaluate all three watermarking methods on `meta-llama/Meta-Llama-3-70B-Instruct`. Due to the inference cost of the 70B model, we use 96 samples from the MMW Story dataset and consider generation lengths of 150 and 250 tokens. We otherwise follow the same detection protocol and watermark-strength settings as in the main experiments. As shown in Table 6, weakening the watermark consistently improves or matches detection performance on the 70B backbone.

D.4 Multi-seed Results

To assess whether the gains from weaker watermarking are robust to randomness in generation

Table 6: Watermark detection performance with `meta-llama/Meta-Llama-3-70B-Instruct` on the MMW Story dataset (96 samples) under different token budgets. We report TPR at fixed FPR thresholds (0.1%, 0.01%, 0.001%) and the median detection p-value across samples. Lower p-values and higher TPR indicate stronger detectability.

Method	150 tokens				250 tokens			
	TPR@FPR			Median	TPR@FPR			Median
	0.1%↑	0.01%↑	0.001%↑	p-value↓	0.1%↑	0.01%↑	0.001%↑	p-value↓
ENS-DiPmark (α=0.5)	13%	4%	1%	7.71e-2	28%	**17%**	6%	1.22e-2
ENS-DiPmark (α=0.4)	**17%**	**8%**	**2%**	**3.81e-2**	**34%**	**17%**	**10%**	**6.62e-3**
SynthID (λ=1)	51%	37%	20%	7.32e-4	76%	58%	44%	3.08e-5
SynthID (λ=0.8)	**62%**	**41%**	**21%**	**2.29e-4**	**81%**	**65%**	**50%**	**9.33e-6**
ENS-MCMark (λ=1)	54%	36%	16%	6.74e-4	67%	58%	45%	**1.90e-5**
ENS-MCMark (λ=0.8)	**56%**	**38%**	**25%**	**5.06e-4**	**72%**	**60%**	**46%**	**1.90e-5**

Table 7: Watermark detection performance across five random seeds (42–46) using Llama-3.2-3B-Instruct with a generation length of 150 tokens. We report the mean and standard deviation of TPR at fixed FPR thresholds (0.1%, 0.01%, and 0.001%). An asterisk (*) denotes statistical significance.

Watermark	**TPR@FPR 0.1%**	**TPR@FPR 0.01%**	**TPR@FPR 0.001%**
ENS-DiPmark ($\alpha = 0.5$)	$51.7 \pm 6.5\%$	$42.0 \pm 6.0\%$	$34.5 \pm 5.6\%$
ENS-DiPmark ($\alpha = 0.4$)	$\mathbf{53.7 \pm 7.4\%}^*$	$\mathbf{44.5 \pm 6.7\%}^*$	$\mathbf{36.5 \pm 5.9\%}^*$
SynthID ($\lambda = 1$)	$75.7 \pm 4.1\%$	$68.0 \pm 4.4\%$	$60.9 \pm 5.6\%$
SynthID ($\lambda = 0.8$)	$\mathbf{77.7 \pm 4.4\%}^*$	$\mathbf{71.0 \pm 6.0\%}^*$	$\mathbf{64.3 \pm 6.3\%}^*$
ENS-MCMark ($\lambda = 1$)	$77.2 \pm 0.8\%$	$70.6 \pm 1.5\%$	$63.5 \pm 2.1\%$
ENS-MCMark ($\lambda = 0.8$)	$\mathbf{79.0 \pm 1.1\%}^*$	$\mathbf{72.4 \pm 1.1\%}^*$	$\mathbf{66.3 \pm 1.9\%}^*$

and watermarking, we repeat the 150-token Llama-3.2-3B-Instruct experiment with five random seeds (42–46). Table 7 reports the mean and standard deviation of TPR across seeds at the three FPR thresholds used throughout the paper.

E Comparison with SynthID

We clarify our contributions in the context of SynthID (Dathathri et al., 2024). While they provide insights into a specific case, our work generalizes these phenomena into a universal theory applicable to all distortion-free ensemble watermarks.

Specifically, the scope of Dathathri et al. (2024) is limited to their proposed Tournament Sampling method (SynthID-Text). Their analytical findings rely on complex, method-specific derivations that are focused exclusively on collision entropy. In contrast, our framework applies broadly to all distortion-free ensemble watermarks. We provide a generalized proof that accommodates all concave entropies, rather than just collision entropy.

Furthermore, to analyze detectability, we introduce the concept of the expected green ratio. Using this concept, we theoretically demonstrate that the expected per-layer detectability will inevitably drop for all existing distortion-free ensemble watermarks because their expected green ratios are concave.

Finally, our work goes beyond theoretical analysis by offering a practical output. We propose a universal, applicable method to construct weaker distortion-free ensemble watermarks (as defined in Eq. 14). We empirically show that this approach yields significant improvements in both long-term detectability and robustness, a practical application not addressed in Dathathri et al. (2024).