Reducing False Positives in Watermark Detection through Entropy-Based Filtering

Anonymous ACL submission

Abstract

With current watermarking systems for LLMs, false positives are statistically rare, but can have severe consequences when individuals are falsely accused of unauthorized AI use. We investigate the popular KGW approach to watermarking and find that low-entropy tokens tend to have less watermark signal. By filtering out low-entropy tokens and adjusting the decision threshold, we substantially reduce false positives without sacrificing detection sensitivity. Specifically, experiments on open-domain news with LLaMA 3.1–8B show a Pareto improvement, reducing false positives by a factor of 5.4 and false negatives by 3%. At a stricter threshold, the method reduces false positives by a factor of 7.7 with only 4% more false negatives. Our insights into the distribution of the watermark signal across entropy levels could further help in recognizing false positives.

1 Introduction

Watermarking has been proposed as a strategy to mitigate potential misuse of large language models (LLMs) (Kamaruddin et al., 2018; Zellers et al., 2020; Weidinger et al., 2021; Touvron et al., 2023). Among existing approaches, the KGW framework (Kirchenbauer et al., 2024a) has emerged as one of the most popular baselines. It biases text generation toward a pseudorandom subset of the vocabulary so that outputs contain more "green" tokens than expected by chance, enabling detection via a z-score test.

Building on KGW, much of the subsequent research has focused on robustness to paraphrasing and editing attacks (Zhao et al., 2023; He et al., 2024; Kirchenbauer et al., 2024b; Zhang and Koushanfar, 2024). The risk of false positives—misclassifying genuine human text—remains underexplored. Although statistically rare (Fernandez et al., 2023; Kirchenbauer et al., 2024a), such errors pose a critical challenge in high-stakes domains such as education, journalism, and law, where even a single error can have severe consequences (Solaiman et al., 2019). This paper addresses this gap by analyzing the sources of false positives and proposing methods to reduce them without sacrificing detection sensitivity.

Prior work has noted repetitions as a source of false positives for watermark detectors (Kirchenbauer et al., 2024a). We extend this analysis to broader cases, showing that predictable tokens, such as named entities, function words, and conventional phrases, often arise from context rather than watermarking. These tokens can also bias detection, motivating our entropy-based filtering approach.

Our main contributions are:

- We demonstrate that entropy-based filtering, previously studied mainly for code generation, is also effective for natural language watermarking.
- We show that, when combined with stricter decision threshold (i.e., z_{th}), entropy-based filtering effectively reduces false positives in natural language watermarking—a connection not established in prior work.

2 Related Work

The KGW framework (Kirchenbauer et al., 2024a) establishes a widely used baseline for watermarking that underlies many subsequent studies. For example, Codable Text Watermarking for LLMs (CTWL) increases payload capacity by partitioning the vocabulary into multiple subsets (Wang et al., 2024). Selective Watermarking via Entropy Thresholding (SWEET) applies watermarking selectively in high-entropy contexts to preserve fluency in code generation (Lee et al., 2024), while Entropy-based Watermarking Detection (EWD) reweights tokens by entropy during detection and reports improved

recall on structured outputs such as code (Lu et al., 2024). These approaches highlight the usefulness of structural modifications and entropy signals but primarily emphasize capacity, robustness, or recall rather than systematic false positive control in open-domain natural language. This limited attention likely stems from the fact that the KGW framework already addressed one source of false positives (repetitions), proposed a simple mitigation by ignoring repeated bigrams[1], and reported an extremely low false positive rate (FPR).

Nevertheless, Fernandez et al. (2023) show that the theoretical FPR guaranteed by KGW under independence assumptions does not hold in practice, as natural language exhibits strong dependencies that lead to much higher realized error rates. They further propose a modular framework with statistical calibration and output smoothing to mitigate this discrepancy. Robust and Imperceptible Watermarking (RIW) (Ren et al., 2024) likewise targets FPR reduction, combining softer embedding with a voted z-test for more conservative error control. These efforts reflect a shift toward watermarking methods that are not only robust but also cautious in deployment, highlighting the need to better understand the sources of false positives and to design approaches that reduce them more reliably. Our study extends this line of work by introducing entropy-based filtering within the KGW framework, specifically aiming to reduce false positives in open-domain natural language watermarking.

3 Method

3.1 Baseline: KGW Watermark Detection

We adopt the KGW framework (Kirchenbauer et al., 2024a) as baseline. At generation time, the vocabulary is partitioned at each decoding step into a *green list* G and a *red list* R, determined pseudorandomly by hashing the preceding h tokens. A fixed fraction γ of the vocabulary is assigned to G, and the logits of green tokens receive a small positive bias δ, making them more likely to be sampled.

At detection time, the same partitioning procedure is applied, and the realized proportion of green tokens $|s|_G$ is compared to the null expectation γ. The deviation is quantified by a z-score:

$$z = \frac{|s|_G - \gamma T}{\sqrt{T\gamma(1-\gamma)}}, \tag{1}$$

where T denotes the number of tokens considered. A text is flagged as watermarked if z exceeds a decision threshold z_{th}. Under independence assumptions, a threshold of $z_{\text{th}} = 4.0$ corresponds to a theoretical FPR of 3×10^{-5}.

3.2 Entropy-based Filtering

Rationale. Tokens predicted with near certainty provide little opportunity for the watermark bias to take effect. Including them in Eq. 1 can therefore distort the detection score. We address this by filtering out such low-entropy tokens before computing z.

Computation. Let $p_v^{(t)}$ be the model's predicted probability of token v at position t, given the preceding context $v_{<t}$. We compute Shannon entropy[2]

$$H_t \;=\; -\sum_{v\in\mathcal{V}} p_v^{(t)} \log p_v^{(t)}. \tag{2}$$

Tokens with $H_t < \tau$ are discarded, yielding an high-entropy subset $\mathcal{S}_\tau = \{t \in [1,T] \mid H_t \geq \tau\}$ with size $T_\tau = |\,\mathcal{S}_\tau|$.

Let $|s_\tau|_G$ be the number of retained tokens that are green; we recompute the statistic on the filtered subset:

$$z_\tau \;=\; \frac{|s_\tau|_G - \gamma T_\tau}{\sqrt{T_\tau\,\gamma(1-\gamma)}}. \tag{3}$$

A text is classified as watermarked if $z_\tau > z_{th}$.

Evaluation. The baseline is the KGW detector without entropy filtering. We evaluate along two dimensions: (i) *cross-threshold* — for a fixed z_{th}, entropy-filtered variants are compared against the unfiltered baseline at the same threshold; (ii) *fixed-τ trajectory* — for a chosen entropy threshold τ, we use $z_{\text{th}} = 4.0$ as the reference point and track performance as z_{th} increases. Figure 1 illustrates both comparison schemes. We report false positive rates (FPRs) and false negative rates (FNRs).

4 Experiments

4.1 Dataset and Preprocessing

We use the *realnewslike* subset of the Colossal Clean Crawled Corpus (C4) (Raffel et al., 2023)

[1] Repeated bigrams (i.e., adjacent token pairs that appear multiple times in the text) are counted only once when computing the watermark statistic, preventing artificially high z-scores from repetitive phrasing. All experiments in this paper follow this setting.

[2] Entropy is computed for each token, but following the detection procedure, only the first occurrence of each bigram is considered; later repetitions are excluded from both the counts and the entropy statistics.

via Hugging Face Datasets[3]. The C4 texts serve as human-written samples (unwatermarked) for detection and as prompts for generating watermarked and unwatermarked completions (see Section 4.2 for details). We discard texts shorter than 190 tokens and truncate longer ones to 200 tokens; after this length control, all texts are normalized with Unicode NFKC[4]. After preprocessing, the final dataset contains 10.8M human-written texts, 20.6K watermarked completions, and 20.5K unwatermarked completions.

4.2 Experimental Setup

LLaMA 3.1–8B (Touvron et al., 2023; Grattafiori et al., 2024) is used for generation and detection, with tokenization handled by its native tokenizer.

Generation. We use the KGW watermarking implementation (Kirchenbauer et al., 2024a) with multinomial sampling (temperature 0.7, max length 250, seed 123). Each completion is prompted with the first 50 tokens of a human-written text. For watermarked completions, we set $\gamma = 0.25$, $\delta = 2.0$, with $h = 1$ (single-token context). A fixed hash key (15485863) is used to ensure reproducibility of the green–red partitioning.

Detection. Building on the KGW detection framework, we incorporate entropy-based filtering and evaluate across $\tau \in \{\text{none}, 0.1, 0.2, \ldots, 1.0, 2.0\}$ and $z_{\text{th}} \in [4.0, 5.0]$ in steps of 0.1 to characterize the FPR–FNR trade-off. The "none" setting indicates detection without entropy filtering, corresponding to the original KGW baseline.

4.3 Results

Figure 1 illustrates Pareto curves for the proposed entropy-based filtering method under varying z_{th} and τ values. The black baseline shows the classical trade-off in watermark detection: higher z_{th} reduces FPRs but increases FNRs, whereas our method shifts this frontier, reducing FPRs while maintaining low FNRs.

Across all z_{th} values, filtering with entropy thresholds in the range $\tau \in [0.1, 1.0]$ consistently reduces FNRs, confirming that low-entropy tokens can obscure watermark signals. For instance, at $z_{\text{th}} = 4.0$, the FNR drops from 3.39% (baseline)

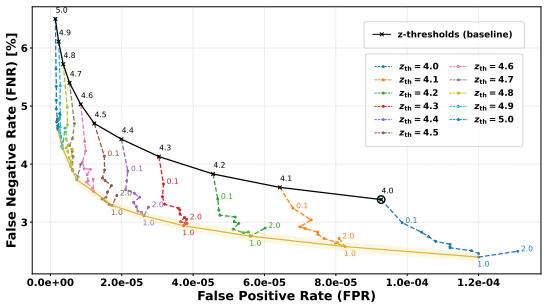


Figure 1: Pareto curves under entropy-based filtering at fixed decision thresholds ($z_{\text{th}} \in [4.0, 5.0]$) and varying entropy thresholds ($\tau \in \{\text{none}, 0.1, 0.2, \ldots, 1.0, 2.0\}$). The black solid line denotes the KGW baseline without entropy filtering, plotted across all z_{th} values. Yellow markers highlight the trajectory for $\tau = 1.0$, shown relative to the baseline reference point at $z_{\text{th}} = 4.0$ on the black curve.

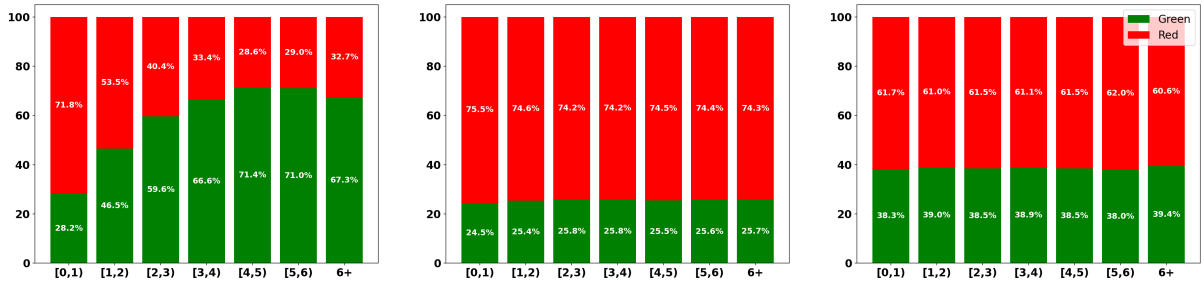
to 2.40% at $\tau = 1.0$, while the FPR slightly increases from 9.27×10^{-5} to 1.20×10^{-4}. This illustrates that entropy filtering effectively reduces false negatives at the cost of a minor increase in false positives. When filtering becomes overly aggressive ($\tau = 2.0$), both FPR and FNR deteriorate, as excessive token removal reduces the effective sample size, weakening the statistical signal for detection. The most substantial gains occur around $\tau \in [0.8, 1.0]$, where FNRs reach their minimum and FPR increases remain marginal.

We next fix τ and analyze performance over varying z_{th}. As shown by the $\tau = 1.0$ trajectory in Figure 1, at $z_{\text{th}} = 4.0$ entropy filtering lowers FNR but slightly raises FPR relative to the KGW baseline. Increasing z_{th} to 4.1–4.5 reverses this FPR rise while keeping FNRs below baseline, marking this range as the most favorable operating point. Specifically, at $z_{\text{th}} = 4.5$, FPR is reduced by a factor of 5.4 (from 9.27×10^{-5} to 1.73×10^{-5}), and FNR decreases by 3% (from 3.39% to 3.29%), illustrating the joint improvement achieved by combining entropy filtering with a slightly stricter detection threshold.

At higher thresholds, the trade-off between FPR and FNR becomes evident. At $z_{\text{th}} = 4.6$, FPR is reduced by a factor of 7.7 (from 9.27×10^{-5} to 1.21×10^{-5}), while FNR increases slightly by about 4% (from 3.39% to 3.53%), preserving a strong balance between sensitivity and specificity. Beyond this point, FPR continues to decrease but with diminishing returns, as the accompanying rise in FNR becomes more pronounced (reaching

[3] https://huggingface.co/datasets/allenai/c4

[4] Following the KGW implementation, NFKC removes ambiguous whitespace and invisible control characters.



(a) (b) (c)

Figure 2: Proportions of green and red tokens across entropy bins in (a) machine-generated watermarked texts, (b) machine-generated unwatermarked texts, and (c) false positives of human-written texts, with the vertical axis indicating token proportions (0–100%).

4.59% at $z_{\text{th}} = 5.0$), indicating that further tightening the threshold yields limited overall benefit.

This pattern generally holds across entropy thresholds, though the effect is most pronounced for $\tau \in [0.8, 1.0]$, where the trajectories form the lower envelope of the Pareto frontier in Figure 1. In this range, FPRs steadily decrease as z_{th} increases, while FNRs remain low up to about $z_{\text{th}} = 4.5$–4.6. Overall, entropy filtering combined with stricter detection thresholds yields Pareto improvements, reducing FPRs while keeping FNRs low.

4.4 Analysis

Entropy-based filtering reveals how watermark signals concentrate in high-uncertainty regions. In watermarked texts (Figure 2a), the green-token proportion rises with entropy and peaks at over 70% in the bin $[4, 5)$, whereas the low-entropy bin $[0, 1)$ is dominated by red tokens (over 70%). This pattern indicates that watermark bias manifests primarily when the model's next-token prediction is uncertain, while near-deterministic tokens dilute the signal. Filtering out such tokens thus strengthens detection and reduces false negatives.

In contrast, unwatermarked outputs[5] (Figure 2b) maintain a stable green proportion around the null expectation $\gamma = 0.25$ across all entropy bins, reflecting the absence of watermark bias and confirming that the correlation observed in panel (a) stems from watermarking rather than model uncertainty.

Human-written false positives (Figure 2c) exhibit a uniformly elevated green proportion of about 38–39% across entropy bins. Hence, filtering out low-entropy tokens does not substantially alter the overall red–green balance, while the smaller sample size amplifies the variance of the z-statistic. The resulting fluctuations can push some cases below or above the decision boundary, producing mixed FPR effects. Empirically, pairing entropy filtering with a slightly higher z_{th} stabilizes detection and further suppresses false positives, yielding an overall FPR reduction.

Beyond these statistical effects, the entropy distribution itself distinguishes human from watermarked texts. True watermarked texts show a clear entropy-dependent contrast in red–green token ratios, whereas human texts remain nearly uniform across bins. This flat profile reflects corpus-level regularities rather than watermark bias and serves as a diagnostic cue for identifying false positives, thus deepening our understanding of false positive behavior and reinforcing the complementary role of entropy-based analysis alongside the standard z-score test.

5 Conclusion

This work addressed the problem of false positives in watermark detection. Our key finding is that entropy filtering, when paired with a stricter decision threshold, is an effective way of improving the trade-off between false positives and false negatives in natural language domains such as news. The entropy profile further provides insight into how watermark signals and human text differ across uncertainty levels, thereby offering a basis for more reliable identification of false positives.

Limitations

Entropy statistics vary across models and domains, so fixed entropy and decision thresholds may not transfer reliably. In this study, we select τ and z_{th} from predefined grids on LLaMA 3.1–8B, and

[5] Both machine-generated unwatermarked texts and correctly classified human-written texts exhibit similar distributions, as neither contains watermark bias.

other models or tokenizers may require retuning. Future work should explore adaptive or model-agnostic strategies that generalize more broadly.

References

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. 2023. Three Bricks to Consolidate Watermarks for Large Language Models. *Preprint*, arXiv:2308.00113.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Zhiwei He, Binglin Zhou, Hongkun Hao, Aiwei Liu, Xing Wang, Zhaopeng Tu, Zhuosheng Zhang, and Rui Wang. 2024. Can Watermarks Survive Translation? On the Cross-lingual Consistency of Text Watermark for Large Language Models. *Preprint*, arXiv:2402.14007.

Nurul Shamimi Kamaruddin, Amirrudin Kamsin, Lip Yee Por, and Hameedur Rahman. 2018. A Review of Text Watermarking: Theory, Methods, and Applications. *IEEE Access*, 6:8011–8028.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2024a. A Watermark for Large Language Models. *Preprint*, arXiv:2301.10226.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. 2024b. On the Reliability of Watermarks for Large Language Models. *Preprint*, arXiv:2306.04634.

Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. 2024. Who Wrote this Code? Watermarking for Code Generation. *Preprint*, arXiv:2305.15060.

Yijian Lu, Aiwei Liu, Dianzhi Yu, Jingjing Li, and Irwin King. 2024. An Entropy-based Text Watermarking Detection Method. *Preprint*, arXiv:2403.13485.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Preprint*, arXiv:1910.10683.

Yubing Ren, Ping Guo, Yanan Cao, and Wei Ma. 2024. Subtle Signatures, Strong Shields: Advancing Robust and Imperceptible Watermarking in Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5508–5519, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release Strategies and the Social Impacts of Language Models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

Lean Wang, Wenkai Yang, Deli Chen, Hao Zhou, Yankai Lin, Fandong Meng, Jie Zhou, and Xu Sun. 2024. Towards Codable Watermarking for Injecting Multi-bits Information to LLMs. *Preprint*, arXiv:2307.15992.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, and 4 others. 2021. Ethical and social risks of harm from Language Models. *Preprint*, arXiv:2112.04359.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2020. Defending Against Neural Fake News. *Preprint*, arXiv:1905.12616.

Ruisi Zhang and Farinaz Koushanfar. 2024. Watermarking Large Language Models and the Generated Content: Opportunities and Challenges. *Preprint*, arXiv:2410.19096.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2023. Provable Robust Watermarking for AI-Generated Text. *Preprint*, arXiv:2306.17439.