MARKTUNE: Advancing the Quality-Detectability Pareto Frontier of Open-Weight LM Watermarking

Yizhou Zhao
University of Pennsylvania
yzzhao@sas.upenn.edu

Steven Wu
Carnegie Mellon University
zstevenwu@cmu.edu

Adam Block
Columbia University
abb2190@columbia.edu

Abstract

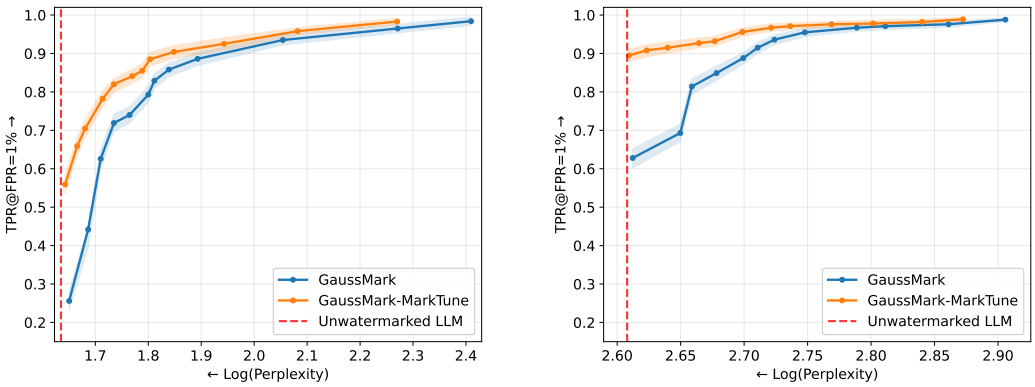
Open-weight language models raise acute challenges for watermarking because inference-time interventions cannot be enforced once model weights are public. Existing methods, such as the recently proposed GAUSSMARK, typically involve subtly modifying model weights. While such schemes demonstrate that imperceptible perturbations can yield detectable signals, they require computationally intensive parameter searches and achieve only limited progress along the quality-detectability frontier. We introduce MARKTUNE, a theoretically principled on-policy fine-tuning framework that treats watermark detectability as a reward signal while regularizing against degradation in text quality. We instantiate our approach with GAUSSMARK as a base watermarking scheme and demonstrate that MARKTUNE consistently improves the quality-detectability trade-off over vanilla GAUSSMARK by adapting non-watermarked weights to maintain generation quality. Empirically, we show that MARKTUNE consistently advances the quality-detectability Pareto frontier: it improves true positive rates under fixed false positive thresholds, restores perplexity and benchmark accuracy to near-unwatermarked levels, and remains robust under paraphrasing and translation attacks. Together, these results establish on-policy fine-tuning as a general strategy for embedding robust, high-quality watermarks into open-weight LMs.

1 Introduction

Open-weight Language Models (LMs) are growing in prevalence due to their rapidly improving capabilities [1, 2, 3]. As open-weight models continue to be deployed, they raise significant concerns about potential misuse on top of the pre-existing societal impacts introduced by closed-weight models. As such, it is critical to develop techniques to ensure appropriate usage that are effective on open-weight models and are sufficiently practical so as to be widely adopted. In this work, we focus on the specific task of *watermarking* LM output, i.e., introducing an almost imperceptible signal into generated text that, when given access to a secret key, can be reliably detected in a statistically valid manner. Watermarking is critical to establish trust that a given piece of text is or is not generated by an LM, which is a necessary prerequisite in a number of societal applications, including academic integrity [4, 5, 6], misinformation mitigation [7, 8, 9], and intellectual property protection [10, 11, 12].

Previous work has posed watermarking as a statistical hypothesis testing problem [13, 14, 15], where a joint distribution is assumed over the text and some watermarking key: in the null hypothesis, the key and text are independent (meaning the text is unwatermarked), while in the alternative hypothesis, the key and text have some statistically detectable relation (meaning the text is watermarked). The goal of a watermarking scheme, then, is to design a mechanism for generating text given a key such that the null and alternative hypotheses can be reliably distinguished, subject to quality constraints on the generated text itself. These quality constraints are often formalized as strict, information-theoretic notions of non-distortion [16, 17, 18] (e.g., the marginal distributions of watermarked and unwatermarked text should be close in total variation distance). In order to satisfy these stringent

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Temperature=0.5, length=200 tokens.

(b) Temperature=0.7, length=200 tokens.

Figure 1: Quality-detectability trade-off of GAUSSMARK and GAUSSMARK-MARKTUNE. See Section 4 for detailed experimental settings.

guarantees while maintaining high detectability, many current approaches to watermarking LMs involve interventions at *inference time* [16, 17, 19, 20, 21, 22], by subtly changing the sampling itself to introduce a watermark signal. While this approach can be effective when the model is accessed only through a generation API, in the case of open-weight models, the provider has no control over a user's generation pipeline and, as such, cannot guarantee that such a watermark will be present in generated text. This problem motivates the need for watermarking techniques specifically designed for open-weight models, where the watermark is embedded directly into the model weights themselves and thus does not require a user to apply a specific decoding approach. Several distortionary watermarking schemes for open-weight models have been proposed that maintain high text quality in practice [13, 23, 24, 25], suggesting that information-theoretic notions of distortion can be overly conservative measures of text quality.

One recently introduced watermarking scheme that intervenes at the level of weights instead of during inference is GAUSSMARK [13], which adds a small amount of Gaussian noise to a subset of the weight matrices, subtly shifting the distribution of generated text in a manner detectable when given access to the added Gaussian noise. In [13], the authors demonstrated that if the variance of the added noise is sufficiently small, and the parameters are carefully chosen, then the text distribution can be modified so as to achieve nontrivial detectability with no loss of text quality. Moreover, [26] demonstrated that GAUSSMARK is at least somewhat robust to a number of simple training-time attacks that a user may apply in an attempt to remove the watermark from the weights of the model. Taken together, these results suggest that GAUSSMARK is a promising approach, but it suffers from the fact that the careful tradeoff between quality and detectability requires a computationally extensive search over parameters and variances so as to find a good set of watermarking hyperparameters. Furthermore, it is not at all clear how close to the Pareto frontier of quality and detectability GAUSSMARK is, and whether or not it is possible to improve upon this tradeoff. We thus ask—*Can we design a watermarking scheme for open-weight LMs that preserves text quality while simultaneously being highly detectable without requiring such a computationally intensive search over parameters?*

Our Contribution. In this work, we answer the above question in the affirmative by proposing MARKTUNE, a novel, theoretically principled, on-policy fine-tuning framework for embedding weight-editing watermarks into open-weight LMs. The core idea is quite simple: turn the watermarking detection into a reward to be optimized during fine-tuning, while simultaneously regularizing the model to maintain high text quality. This procedure allows the model to adapt to perturbations in the weights that would normally harm the quality of generations in a way that preserves the watermark signal. Our framework has the benefit of preserving statistical validity of detection, in the sense that a resulting watermark test maintains whatever statistically rigorous guarantees on false positives the underlying scheme offers. We operationalize our framework with GAUSSMARK as a base watermarking scheme and conduct a number of empirical evaluations and ablations to demonstrate the superiority of our approach over vanilla GAUSSMARK.

Related Work. LM text watermarking schemes can be broadly categorized into two families: *inference-time* watermarking and *model-embedded* watermarking. Distortionary inference-time schemes modify the sampling process—for example, by biasing next-token sampling toward a

partitioned "green list" [19, 27]. Although these methods provide statistical guarantees, they introduce noticeable distortion in generated text and are vulnerable to paraphrasing attacks [28, 29]. In contrast, nondistortionary inference-time schemes embed watermark signals by influencing the pseudorandom number generator used in next-token sampling while preserving the original distribution. For instance, [30] and [16] draw independent pseudorandom variables and generate tokens using deterministic decoders based on the Gumbel-max trick and inverse transform sampling. Similarly, [22] and [21] propose unbiased variants of the KGW watermark [19] by introducing decoding algorithms based on maximal coupling and reweighting strategies, respectively. However, these approaches are not yet ready for large-scale LM deployment due to their generation latency [13] and the fact that they can affect text quality [31]. More recently, [20] introduced a tournament-based watermarking, which achieves high detection power with minimal latency. Yet, maintaining text quality in this setting requires storage that scales linearly with the number of generated tokens, making it impractical for large production systems.

Model-embedded watermarking can be divided into two categories: training-based schemes [23, 24] and weight-editing schemes [13, 25]. These approaches embed the watermark signal directly into model weights, making them naturally suitable for open-weight LMs while incurring neither generation latency nor additional storage overhead. However, training-based schemes remain limited in their ability to generalize across tasks [24] and lack rigorous guarantees on the statistical validity of detection [23]. Weight-editing schemes, in contrast, either require modifications to standard model architectures [25] or suffer from computationally intensive parameter searches and limited advancement in balancing text quality with detection performance [13].

2 Preliminaries

A language model is any conditional distribution mapping a *prompt* $x \in \mathcal{X}$ (the space of prompts) to a distribution over responses $y \in \mathcal{Y}$ (the space of responses), i.e. a function $p : \mathcal{X} \to \Delta(\mathcal{Y})$. As is common in language modeling, we will generally consider *autoregressive generation*, where there is some vocabulary set $\mathcal{V}$ and both $\mathcal{X}$ and $\mathcal{Y}$ are subsets of $\mathcal{V}^*$. In this case, the model generates a response one token at a time by sampling $y_1 \sim p(\cdot|x)$, then $y_t \sim p(\cdot|x, y_1, \ldots, y_{t-1})$ and concatenating the output tokens to form a response. As we are chiefly concerned with transformer instantiations of language models, we generally parameterize the model by some set of weights $\Theta \subset \mathbb{R}^d$ and write p_θ for the resulting model. Typically, in the case of transformers, $\theta \in \Theta$ can be thought of as the concatenation of a large number of high dimensional matrices, one for each layer of the transformer.

Hypothesis Testing. As in [13, 14, 15], we formalize the notion of watermarking as a statistical hypothesis testing problem. Recall that a hypothesis testing problem consists of an observation space $\Xi \times \mathcal{Y}$ and two disjoint collections of distributions on the observation space, $\mathbf{H_0}$ and $\mathbf{H_A}$. A test is a (possibly randomized) function $\phi : \Xi \times \mathcal{Y} \to \{0, 1\}$, where $\phi(\xi, y) = 1$ indicates that the observation (ξ, y) provides sufficient evidence to suggest that it was not sampled from any distribution in $\mathbf{H_0}$. The test is said to have level α if the false positive rate, the probability that $\phi = 1$ even when (ξ, y) is sampled from an element of the null hypothesis, is at most α. The power of the test, $1 - \beta$, is the probability that $\phi = 1$ when (ξ, y) is truly sampled from an element of the alternative hypothesis. Clearly we wish to have a test with both α and β as small as possible.

Weight-editing Watermarking. Watermarking occurs in two phases: generation and detection. Formally, we suppose that there is a watermarking key space Ξ and distribution ρ. In the generation process, the generator samples $\xi \sim \rho$, chooses some $\theta(\xi)$ and samples $y \sim p_{\theta(\xi)}(\cdot|x)$. Detection is phrased as a hypothesis test, where $\mathbf{H_0} = \{\rho \otimes q | q \in \Delta(\mathcal{Y})\}$ the set of distributions where the key ξ and text y are independent, and $\mathbf{H_A}$ is precisely the distribution induced by the generating process[1].

GaussMark. We instantiate our framework with GAUSSMARK [13], a recently proposed weight-editing watermarking scheme that we briefly review here. Given a language model $p_\theta : \mathcal{X} \to \Delta(\mathcal{Y})$, GAUSSMARK partitions the parameter as $\theta = (\theta_{\text{wm}}, \theta_0)$, where θ_{wm} (with dimension d_r) is the subset of model weights modified to embed the watermark, and θ_0 the remaining weights. The base model p_θ is stored as a reference model $q_{\theta'}$ for later detection. To embed a watermark, GAUSSMARK samples the key $\xi_\sigma \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_r})$ and obtains watermarked model $p_{\theta(\xi_\sigma)}$ with $\theta(\xi_\sigma) = (\theta_{\text{wm}} + \xi_\sigma, \theta_0)$, i.e., it perturbs the selected weights with a small amount of Gaussian noise and leaves the others

[1] Observe that as stated, the detector has access to the prompt x used to generate y. In practice, this is of course not the case and our empirical results do not rely on this access.

unchanged. To detect the watermark, GAUSSMARK uses the following test statistic:

$$\psi(y, \xi_\sigma \mid x) = \frac{\langle \xi_\sigma, \nabla_{\theta_{\mathrm{wm}}} \log q_{\theta'}(y \mid x) \rangle}{\sigma \|\nabla_{\theta_{\mathrm{wm}}} \log q_{\theta'}(y \mid x)\|_2}. \tag{1}$$

Intuitively, this statistic measures the alignment between the secret key ξ_σ and the gradient of the reference model with respect to the watermarked weights. Under $\mathbf{H_0}$, ξ_σ is independent of the text y, so $\psi(y, \xi_\sigma \mid x)$ follows a standard normal distribution and a test of level α can be constructed by thresholding the statistic at the inverse Gaussian CDF (denoted by Φ^{-1}) at $1 - \alpha$.

3 Our Method

In this work, we aim to advance the quality-detectability Pareto frontier of weight-editing watermarking, which modifies a subset of model weights to embed watermark signal. Unlike inference-time watermarking, these approaches are applicable in open-weight settings but cannot satisfy information-theoretic guarantees of distortion-freeness. Rather, they introduce a more opaque form of distortion, the impact of which on generation quality remains difficult to quantify. The following proposition characterizes an upper bound on the total variation (TV) distance induced by GAUSSMARK.

Proposition 1. *Given a base language model p_θ with $\theta = (\theta_{\mathrm{wm}}, \theta_0)$ and a sampled Gaussian noise $\xi_\sigma \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_r})$, let $\theta(\xi_\sigma) = (\theta_{\mathrm{wm}} + \xi_\sigma, \theta_0)$ and $p_{\theta(\xi_\sigma)}$ denotes the watermarked model with selected subset of weights perturbed by ξ_σ. Then the induced total variation (TV) distance from the base model can be bounded as $\mathbb{E}_{\xi_\sigma}\left[\sup_{x \in \mathcal{X}} \left\| p_{\theta(\xi_\sigma)}(\cdot \mid x) - p_\theta(\cdot \mid x) \right\|_{\mathrm{TV}}\right] \lesssim \sigma\sqrt{d_r}$.*

Because TV distance characterizes the difficulty of hypothesis testing, GAUSSMARK's detection power scales approximately with this TV distance. In other words, if TV distance from the base model measures the watermark distortion on generated text quality, it seems that the quality-detectability trade-off cannot be improved: increasing power (via larger σ or a higher-dimensional perturbation subspace d_r) inflates an upper bound on the TV distance-based distributional distortion.

Nevertheless, we argue that such pessimism is overstated for two reasons. First, the base model p_θ should not be regarded as an oracle that perfectly characterizes high-quality text. Consequently, *closeness in TV distance is not a necessary condition for achieving high-quality generation*. TV distance is an especially stringent metric, as it upper-bounds worst-case deviations across all possible events, which is far stricter than what is required for human-perceived quality. Second, modern LMs are heavily over-parameterized and exhibit wide, flat optimization basins. For fixed watermarked weights $\theta^\star_{\mathrm{wm}}$ (e.g., $\theta^\star_{\mathrm{wm}} = \theta_{\mathrm{wm}} + \xi_\sigma$ in GAUSSMARK) within a small subspace, there may exist alternative configurations of the remaining weights $\theta^\star_0$ that preserve a significant watermark signal while retaining generation quality by exploiting the large "null space" orthogonal to the watermark. Our framework, MARKTUNE, is designed to exploit both of these observations.

Initial Idea: Supervised Fine-Tuning (SFT). A natural strategy to recover the generation quality is fine-tuning the watermarked model on a labeled dataset. To enhance out-of-distribution (OOD) generalization, we freeze the watermarked weights θ_{wm} and only update the remaining weights θ_0. Without freezing, a strong watermark signal may arise merely from memorization of prompts in the training corpus and fail to generalize to unseen prompt distributions. However, we observe that watermark signal decays rapidly during training. The underlying mechanism behind this phenomenon is co-adaptation: since the supervised cross-entropy (CE) loss imposes no constraint on preserving the watermark signal, the unfrozen weights may adapt in directions that anti-correlate with the fixed watermark perturbation in representation space, thereby accelerating loss minimization but potentially weakening the watermark signal—even though the watermarked weights themselves remain unchanged. See ablation study in Appendix E.3 for an empirical justification.

Refinement: On-Policy Fine-Tuning with Dual Objectives (MARKTUNE). Inspired by the reinforcement learning with verifiable reward (RLVR), we treat the watermark signal as a reward function while incorporating supervised CE loss as a regularization term. Specifically, given a labeled corpus $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^N$, which may be constructed for instruction following or causal language modeling (CLM) objective, and a watermarked model $p_{(\theta^\star_{\mathrm{wm}}, \theta_0)}$ obtained by embedding watermark using a key ξ, we optimize the objective:

$$\max_{\theta_0} \; \mathbb{E}_{(x,y^\star)\sim\mathcal{D},\, y\sim p_{(\theta^\star_{\mathrm{wm}},\theta_0)}(\cdot|x)}\big[\mathcal{R}_{\mathrm{wm}}(x, y; \xi)\big] \; - \lambda \mathcal{L}_{\mathrm{ce}}(\theta^\star_{\mathrm{wm}}, \theta_0;\, x, y^\star), \tag{2}$$

where $\mathcal{R}_{\mathrm{wm}}$ denotes the watermark signal reward and $\mathcal{L}_{\mathrm{ce}}$ is cross-entropy loss. We let $\lambda > 0$ be a hyperparameter that balances watermark reward against fidelity to ground-truth labels. By rewarding

generated samples $y \sim p_{(\theta^\star_{\text{wm}},\theta_0)}(\cdot \mid x)$ that exhibit stronger watermark signals, while simultaneously penalizing deviations from the labeled response via CE loss, this dual-objective, on-policy fine-tuning framework advances the Pareto frontier *without sacrificing statistical rigor*.

Application to GaussMark. We apply our proposed framework to GAUSSMARK. A natural approach would be to replace the watermark reward $\mathcal{R}_{\text{wm}}$ in Eq.(2) with the test statistic in Eq.(1). However, in practice, a model provider or third-party auditor attempting to verify whether a suspect text y is watermarked typically has white-box access to the base model $q_{\theta'}$, but not to the prompt x that was originally used to generate y. This motivates a more practical formulation of the watermark reward and test statistic for GAUSSMARK:

$$\mathcal{R}_{\text{wm}}(y;\xi_\sigma) = \psi(y;\xi_\sigma) = \frac{\langle \xi_\sigma, \nabla_{\theta_{\text{wm}}} \log q_{\theta'}(y)\rangle}{\sigma \|\nabla_{\theta_{\text{wm}}} \log q_{\theta'}(y)\|_2}. \tag{3}$$

By substituting (3) into the objective in (2), we adapt GAUSSMARK to our on-policy fine-tuning framework and obtain GAUSSMARK-MARKTUNE. All algorithm implementation details are provided in Appendix A. For watermark detection, we inherit the procedure in GAUSSMARK. Since the generated text y remains independent of ξ_σ under $\mathbf{H_0}$, Proposition 2 provides rigorous statistical guarantees on the controllability of the false positive rate.

Proposition 2. *Let $\alpha \in (0,1)$, and $\tau_\alpha := \Phi^{-1}(1-\alpha)$ where Φ is the CDF of the standard normal distribution. Then, for any $y \in \mathcal{Y}$, the test $\mathbb{I}\Big\{\frac{\langle \xi_\sigma, \nabla_{\theta_{\text{wm}}} \log q_{\theta'}(y)\rangle}{\sigma \|\nabla_{\theta_{\text{wm}}} \log q_{\theta'}(y)\|_2} \geq \tau_\alpha\Big\}$ has level α.*

Informal Analysis. Intuition behind the success of MARKTUNE over GAUSSMARK can be found in the simple setting, where the last layer of a model is watermarked. In this case, the earlier layers "featurize" text so the model predicts $p_\theta(y \mid x) \propto \exp\{\langle \theta_{\text{wm}}, \phi_{\theta_0}(x,y)\rangle\}$, where $\phi_{\theta_0}(x,y)$ are features produced by the non-watermarked layers. GAUSSMARK replaces this distribution by $p_\theta(y \mid x) \propto \exp\{\langle \theta_{\text{wm}} + \xi_\sigma, \phi_{\theta_0}(x,y)\rangle\}$, yielding a detectable but potentially misaligned output. The core mechanism of MARKTUNE is to only adjust θ_0 so as to *reshape the features* and recover high-quality generation under the frozen watermarked weights $\theta_{\text{wm}} + \xi_\sigma$. As illustrated in Appendix C, MARKTUNE searches for a new $\theta_0^\star$ such that the distribution $p_{(\theta_{\text{wm}}+\xi_\sigma,\, \theta_0^\star)}(y \mid x) \propto \exp\{\langle \theta_{\text{wm}} + \xi_\sigma, \phi_{\theta_0^\star}(x,y)\rangle\}$ moves *closer* (in cross-entropy or KL divergence) to a high-quality target distribution $p^\star(y \mid x)$, while remaining *far* from the original unwatermarked model $p_\theta(y \mid x)$ along the watermark-sensitive direction. Intuitively, the non-watermarked layers learn to *absorb* the bias introduced by ξ_σ, thereby restoring generation quality, while detectability is retained, ensuring that the model remains statistically distinguishable from the unwatermarked base model. This leads to a uniformly improved quality-detectability Pareto frontier.

4 Experiments

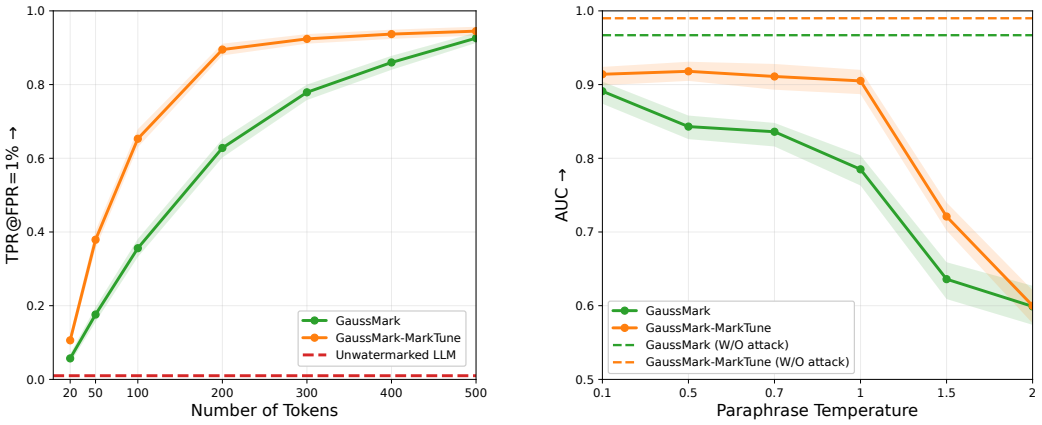
Experimental Setup. Unless otherwise noted, we use GPT-2 with 124M parameters [32] to generate responses of 200 tokens at a temperature of 0.7. In line with prior watermarking work [13, 16, 19], we evaluate watermark performance using 1K prompts from the realnewslike split of the C4 dataset [33]. **For all experiments, watermark detection is conducted using only the generated responses, aligning with practical deployment scenarios.** For fine-tuning, we use OpenWebText [34] as the training and validation corpus, following the implementation of `nanoGPT`.[2] To evaluate watermark detectability, we report the true positive rate (TPR) for a fixed false positive rate (FPR) of 1% as well as ROC curves and the area thereunder (AUC). To measure text quality, we examine (1) perplexity (PPL) of generations using OPT-2.7B [35] as a larger oracle language model, (2) validation loss on OpenWebText, and (3) the LAMBADA benchmark [36] preprocessed by OpenAI, which is designed to evaluate GPT-2's ability to perform long-range text understanding. The evaluation metrics for LAMBADA benchmark include accuracy (ACC) and perplexity (PPL). For evaluation on robustness against paraphrasing, we use the `T5_Paraphrase_Paws` model [37] and control its strength by adjusting the sampling temperature. Implementation details are provided in Appendix D.

Quality-Detectability Trade-off. Figure 1 presents the detection performance and the text quality of distortionary vanilla GAUSSMARK and GAUSSMARK-MARKTUNE under different generation temperatures. We plot quality-detectability trade-off curves by adjusting the hyperparameters σ and λ. Compared with vanilla GAUSSMARK, GAUSSMARK-MARKTUNE achieves a more favorable trade-off, yielding substantially higher detection rates for the same distortion on text quality.

[2] https://github.com/karpathy/nanoGPT.

Table 1: Generated text quality across different metrics. Models highlighted in blue exhibit comparable generation quality to the unwatermarked model and are thus selected for later experiments.

Model	PPL↓	Val. Loss↓	LAMBADA (ACC)↑	LAMBADA (PPL)↓
Unwatermarked	13.57 ± 5.58	3.121	$.3093 \pm .0064$	38.43 ± 1.414
GAUSSMARK ($\sigma = 0.04$)	13.63 ± 5.41	3.124	$.3008 \pm .0064$	39.33 ± 1.503
GAUSSMARK ($\sigma = 0.1$)	18.28 ± 7.48	3.305	$.2474 \pm .0060$	70.46 ± 2.817
GAUSSMARK-MARKTUNE ($\sigma = 0.04$)	12.69 ± 5.01	3.086	$.3305 \pm .0066$	35.82 ± 1.386
GAUSSMARK-MARKTUNE ($\sigma = 0.1$)	13.59 ± 5.29	3.118	$.3059 \pm .0064$	37.97 ± 1.423



(a) TPR@FPR=1% as a function of text length.

(b) AUC as a function of paraphrasing temperature.

Figure 2: Watermark detectability and robustness against paraphrasing under minimal distortion.

Generation Performance and Robustness. Table 1 shows the generation quality of models under different watermarking regimes, evaluated using four metrics. We observe that GAUSSMARK-MARKTUNE consistently outperforms vanilla GAUSSMARK under the same hyperparameter σ across all text quality metrics, demonstrating its effectiveness in restoring generation quality. To ensure a fair detectability comparison under minimal distortion, We select σ for each watermarking regime by maximizing detectability while ensuring that the text quality metrics remain comparable to those of the unwatermarked counterpart. Based on the empirical results, we set $\sigma = 0.04$ for vanilla GAUSSMARK and $\sigma = 0.1$ for GAUSSMARK-MARKTUNE in later experiments.

Figure 2 demonstrates that applying our on-policy fine-tuning framework to GAUSSMARK substantially improves detectability and robustness to paraphrasing attacks while preserving generation quality. Specifically, Figure 2a shows that the TPR@FPR=1% of GAUSSMARK-MARKTUNE is uniformly higher than those of vanilla GAUSSMARK, with the largest gains for short sequences (50–200 tokens). This trend is attributed to watermark signal saturation, as evidenced by the flattening of $\|\nabla_{\theta_{\text{wm}}} \log q_{\theta'}(y)\|_2$ reported in [13]. Figure 2b also exhibits holistic improvement. In particular, GAUSSMARK-MARKTUNE remains stable under mild paraphrasing temperatures, indicating strong resilience to lexical variation. Additional empirical results on watermark detectability and robustness, along with a comprehensive ablation study demonstrating the effect different parameter choices have on detectability and quality (cf. Table 2), are provided in Appendix E.

5 Conclusion

We introduced MARKTUNE, a practical and theoretically grounded on-policy fine-tuning framework for weight-editing watermarks. By optimizing a dual objective that combines a watermark signal reward with supervised cross-entropy regularization, MARKTUNE adapts non-watermarked parameters to accommodate fixed watermark edits. Extensive empirical evaluations demonstrate that our approach advances the quality-detectability Pareto frontier of vanilla GAUSSMARK and improves robustness against diverse attacks, while maintaining statistical guarantees and incurring neither generation nor storage overhead. While some empirical work [26] has demonstrated some degree of robustness of GAUSSMARK to finetuning attacks, further investigation is warranted to assess the resilience of MARKTUNE against more sophisticated adaptive adversaries.

References

[1] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[2] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[3] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[4] Chris Stokel-Walker. Ai bot chatgpt writes smart essays-should professors worry? *Nature News*, 2022.

[5] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[6] Silvia Milano, Joshua A McGrane, and Sabina Leonelli. Large language models challenge the future of higher education. *Nature Machine Intelligence*, 5(4):333–334, 2023.

[7] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In *Advances in neural information processing systems*, 2019.

[8] Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1):82, 2020.

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

[10] Claudio Novelli, Federico Casolari, Philipp Hacker, Giorgio Spedicato, and Luciano Floridi. Generative ai in eu law: Liability, privacy, intellectual property, and cybersecurity. *Computer Law & Security Review*, 55:106066, 2024.

[11] Yongqi Jiang, Yansong Gao, Chunyi Zhou, Hongsheng Hu, Anmin Fu, and Willy Susilo. Intellectual property protection for deep learning model and dataset intelligence. *arXiv preprint arXiv:2411.05051*, 2024.

[12] Zhenhua Xu, Xubin Yue, Zhebo Wang, Qichen Liu, Xixiang Zhao, Jingxuan Zhang, Wenjun Zeng, Wengpeng Xing, Dezhang Kong, Changting Lin, et al. Copyright protection for large language models: A survey of methods, challenges, and trends. *arXiv preprint arXiv:2508.11548*, 2025.

[13] Adam Block, Ayush Sekhari, and Alexander Rakhlin. Gaussmark: A practical approach for structural watermarking of language models. In *International Conference on Machine Learning*, 2025.

[14] Baihe Huang, Hanlin Zhu, Banghua Zhu, Kannan Ramchandran, Michael I Jordan, Jason D Lee, and Jiantao Jiao. Towards optimal statistical watermarking. *arXiv preprint arXiv:2312.07930*, 2023.

[15] Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *The Annals of Statistics*, 53(1):322–351, 2025.

[16] Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research*, 2024.

[17] Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. In *International Conference on Learning Representations*, 2024.

[18] Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1125–1139. PMLR, 2024.

[19] 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*, 2023.

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

[21] Yihan Wu, Zhengmian Hu, Hongyang Zhang, and Heng Huang. DipMark: A stealthy, efficient and resilient watermark for large language models. *arXiv preprint arXiv:2310.07710*, 2023.

[22] Yangxinyu Xie, Xiang Li, Tanwi Mallick, Weijie J Su, and Ruixun Zhang. Debiasing watermarks for large language models via maximal coupling. *arXiv preprint arXiv:2411.11203*, 2024.

[23] Xiaojun Xu, Yuanshun Yao, and Yang Liu. Learning to watermark llm-generated text via reinforcement learning. *arXiv preprint arXiv:2403.10553*, 2024.

[24] Chenchen Gu, Xiang Lisa Li, Percy Liang, and Tatsunori Hashimoto. On the learnability of watermarks for language models. In *International Conference on Learning Representations*, 2024.

[25] Miranda Christ, Sam Gunn, Tal Malkin, and Mariana Raykova. Provably robust watermarks for open-source language models. *arXiv preprint arXiv:2410.18861*, 2024.

[26] Thibaud Gloaguen, Nikola Jovanović, Robin Staab, and Martin Vechev. Towards watermarking of open-source llms. *arXiv preprint arXiv:2502.10525*, 2025.

[27] Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for AI-generated text. In *International Conference on Learning Representations*, 2024.

[28] Saksham Rastogi and Danish Pruthi. Revisiting the robustness of watermarking to paraphrasing attacks. *arXiv preprint arXiv:2411.05277*, 2024.

[29] Nikola Jovanović, Robin Staab, and Martin Vechev. Watermark stealing in large language models. In *International Conference on Machine Learning*, 2024.

[30] Scott Aaronson. Watermarking of large language models. https://simons.berkeley.edu/talks/scott-aaronson-ut-austin-openai-2023-08-17, August 2023.

[31] Yihan Wu, Ruibo Chen, Zhengmian Hu, Yanshuo Chen, Junfeng Guo, Hongyang Zhang, and Heng Huang. Distortion-free watermarks are not truly distortion-free under watermark key collisions. *arXiv preprint arXiv:2406.02603*, 2024.

[32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[33] 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.

[34] Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

[35] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[36] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*, 2016.

[37] Sai Vamsi Alisetti. Paraphrase generator with t5. https://huggingface.co/Vamsi/T5_Paraphrase_Paws, 2021.

[38] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[39] Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. Robust detection of watermarks for large language models under human edits. *arXiv preprint arXiv:2411.13868*, 2024.

[40] Zhiwei He, Binglin Zhou, Hongkun Hao, Aiwei Liu, Xing Wang, Zhaopeng Tu, Zhuosheng Zhang, and Rui Wang. Can watermarks survive translation? on the cross-lingual consistency of text watermark for large language models. *arXiv preprint arXiv:2402.14007*, 2024.

[41] Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raúl Vázquez, and Sami Virpioja. Democratizing neural machine translation with opus-mt. *Language Resources and Evaluation*, 58(2):713–755, 2024.

A Practical Implementation of MARKTUNE

To optimize the objective in Eq.(2), we introduce a GRPO-style [38] policy optimization algorithm. This choice is motivated by the fact that Eq.(2) cannot be directly optimized due to its dependence on sampled outputs $y \sim p_{(\theta^\star_{\text{wm}},\theta_0)}$. GRPO is a critic-free, on-policy policy-gradient method that replaces the learned value baseline in PPO with a group-relative baseline computed from multiple samples per prompt, which makes it well-suited for optimizing our objective.

We freeze the watermarked weights $\theta^\star_{\text{wm}}$ and conduct the optimization algorithm. For each prompt from a batch $x \sim \mathcal{D}_b$, we sample a group of responses $\{y_j\}_{j=1}^G \sim p_{(\theta^\star_{\text{wm}},\theta_0)}(\cdot \mid x)$ and compute their watermark signal rewards $\mathbf{r} = \{r_j\}_{j=1}^G$. The advantage for sample j is normalized within the group:

$$\widehat{A}_j = \big(r_j - \text{mean}(\mathbf{r})\big)/\text{std}(\mathbf{r}) \tag{4}$$

The policy update follows a clipped objective based on the importance ratio $\rho_j = \frac{p_{(\theta^\star_{\text{wm}},\theta_0)}(y_j|x)}{p_{(\theta^\star_{\text{wm}},\theta_0^{\text{old}})}(y_j|x)}$, maximizing $\mathbb{E}\Big[\min\big(\rho_j \widehat{A}_j,\ \text{clip}(\rho_j, 1-\epsilon, 1+\epsilon)\widehat{A}_j\big)\Big]$ together with a supervised cross-entropy regularization term. Denote $\theta = (\theta^\star_{\text{wm}}, \theta_0)$, the overall GRPO-style objective is give by

$$\mathcal{J}(\theta) = \mathbb{E}_{(x,y^\star)\sim\mathcal{D},\{y_j\}_{j=1}^G\sim p_\theta(\cdot|x)} \frac{1}{G}\sum_{j=1}^{G} \min\left[\frac{p_\theta(y_j \mid x)}{p_{\theta_{\text{old}}}(y_j \mid x)}\hat{A}_j,\ \text{clip}\left(\frac{p_\theta(y_j \mid x)}{p_{\theta_{\text{old}}}(y_j \mid x)},\ 1-\varepsilon,\ 1+\varepsilon\right)\hat{A}_j\right] - \lambda\mathcal{L}_{\text{ce}}(\theta; x, y^\star). \tag{5}$$

See Algorithm 1 for a high-level implementation of MARKTUNE framework. See Algorithm 2 and Algorithm 3 for the detailed implementation of GRPO-style policy optimization and application of MARKTUNE to GAUSSMARK.

Algorithm 1 MARKTUNE Meta-Algorithm

Input: Language model p_θ with $\theta = (\theta_{\text{wm}}, \theta_0)$, watermark key ξ, weight-editing watermark algorithm $\mathcal{A}(\cdot, \xi)$, labeled corpus $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^N$, CE coefficient λ.
Conduct weight-editing watermarking: $\theta^\star_{\text{wm}} \leftarrow \mathcal{A}(\theta_{\text{wm}}, \xi)$.
Freeze $\theta^\star_{\text{wm}}$ and finetune θ_0 to optimize Eq. (2) with Algorithm 2 to obtain $\theta^\star_0$.
Output: Watermarked weights $(\theta^\star_{\text{wm}}, \theta^\star_0)$.

Algorithm 2 GRPO-style Policy Optimization for MARKTUNE

Input: initial policy model $p_{(\theta^\star_{\text{wm}},\theta_0)}$; watermark signal reward $\mathcal{R}_{\text{wm}}(\cdot,\cdot;\xi_\sigma)$; labeled corpus $\mathcal{D}$; hyperparameters $\varepsilon, \lambda, T_1, T_2$.
Initialize: Freeze $\theta^\star_{\text{wm}}$ and set $p_\theta \leftarrow p_{(\theta^\star_{\text{wm}},\theta_0)}$.
for step $= 1, \ldots, T_1$ **do**
  Sample a batch $\mathcal{D}_b \subset \mathcal{D}$.
  $p_{\theta_{\text{old}}} \leftarrow p_\theta$.
  for each prompt $x \in \mathcal{D}_b$ **do**
    Sample G outputs $\{y_j\}_{j=1}^G \sim p_{\theta_{\text{old}}}(\cdot \mid x)$.
    Compute rewards $\{r_j\}_{j=1}^G$ using $\mathcal{R}_{\text{wm}}$.
    Compute $\hat{A}_j$ for response y_j via group-relative advantage Eq.(4).
  end for
  for iteration $= 1, \ldots, T_2$ **do**
    Update p_θ by maximizing the GRPO-style objective Eq.(5).
  end for
end for
Output: p_θ.

Algorithm 3 Working Pipeline of GAUSSMARK-MARKTUNE

1: **Input:** Language model p_θ with $\theta = (\theta_{\text{wm}}, \theta_0)$, fixed base model $q_{\theta'} = p_\theta$, strength $\sigma > 0$, labeled corpus $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^N$, CE coefficient λ.
2: Sample watermark key $\xi_\sigma \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_r})$.
3: Inject GaussMark perturbation $\theta^\star_{\text{wm}} \leftarrow \theta_{\text{wm}} + \xi_\sigma$.
4: Freeze $\theta^\star_{\text{wm}}$ and conduct GRPO-style policy update Algorithm 2 for θ_0 to obtain $\theta_0^\star$ via optimizing the objective

$$\max_{\theta_0} \mathbb{E}_{(x,y^\star)\sim\mathcal{D}, y\sim p_{(\theta^\star_{\text{wm}},\theta_0)}(\cdot|x)} \left[\frac{\langle \xi_\sigma,\ \nabla_{\theta_{\text{wm}}} \log q_{\theta'}(y) \rangle}{\sigma \|\nabla_{\theta_{\text{wm}}} \log q_{\theta'}(y)\|} \right] - \lambda \mathcal{L}_{\text{ce}}(\theta^\star_{\text{wm}}, \theta_0;\, x, y^\star),$$

5: **Output:** Watermarked parameters $(\theta^\star_{\text{wm}}, \theta_0^\star)$.

B Theorems and Proofs

B.1 Proof of Proposition 1

Proof. Given a prompt x, we denote $p_\theta(\cdot \mid x)$ by p_θ for simplicity. By Pinsker's inequality,

$$\left\| p_{\theta(\xi_\sigma)} - p_\theta \right\|_{\text{TV}} \ \le\ \sqrt{\frac{1}{2} \operatorname{KL}\left(p_{\theta(\xi_\sigma)} \middle\| p_\theta \right)}, \tag{6}$$

where $\xi_\sigma \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_r})$. Let $D(\xi_\sigma) \ :=\ \operatorname{KL}\left(p_{\theta(\xi_\sigma)} \middle\| p_\theta \right) \ge 0$. Since $D(0) = 0$ and its gradient vanishes at the global minimum $\left(\nabla_{\xi_\sigma} D(\xi_\sigma) \mid_{\xi_\sigma = 0} = 0 \right)$, its Taylor expansion around $\xi_\sigma = 0$ begins at second order:

$$D(\xi_\sigma) = \frac{1}{2} \xi_\sigma^\top \mathcal{I}(\theta_{\text{wm}})\, \xi_\sigma + o\big(\|\xi_\sigma\|^2\big), \quad \mathcal{I}(\theta_{\text{wm}}) := \nabla^2_{\xi_\sigma} D(\xi_\sigma)\big|_{\xi_\sigma = 0},$$

where $\mathcal{I}(\theta_{\text{wm}}) \in \mathbb{R}^{d_r \times d_r}$ is the Fisher information matrix. Each diagonal entry $\mathcal{I}_{jj}(\theta_{\text{wm}})$ of $\mathcal{I}(\theta_{\text{wm}})$ represents the Fisher information of j-th component $\theta^{(j)}_{\text{wm}}$ of θ_{wm} and measures how much information a single model response y provides about the specific parameter component $\theta^{(j)}_{\text{wm}}$:

$$\mathcal{I}_{jj}(\theta_{\text{wm}}) = \mathbb{E}_{y\sim p_\theta(\cdot|x)} \left[\left(\frac{\partial \log p_\theta(y \mid x)}{\partial \theta^{(j)}_{\text{wm}}} \right)^2 \right].$$

Then we take the expectation of this approximation with respect to the distribution of ξ_σ:

$$\begin{aligned}
\mathbb{E}[D(\xi_\sigma)] &= \mathbb{E}\left[\frac{1}{2} \xi_\sigma^\top \mathcal{I}(\theta_{\text{wm}})\, \xi_\sigma + o\big(\|\xi_\sigma\|^2\big) \right] \\
&= \frac{1}{2} \mathbb{E}\left[\operatorname{tr}(\mathcal{I}(\theta_{\text{wm}}) \xi_\sigma \xi_\sigma^\top) \right] + o(\sigma^2 d_r) \\
&= \frac{1}{2} \operatorname{tr}\left\{ \mathcal{I}(\theta_{\text{wm}}) \mathbb{E}\left[\xi_\sigma \xi_\sigma^\top \right] \right\} + o(\sigma^2 d_r) \\
&= \frac{\sigma^2}{2} \operatorname{tr}(\mathcal{I}(\theta_{\text{wm}})) + o(\sigma^2 d_r).
\end{aligned}$$

For a well-defined model, there exists a model-dependent constant capturing the local Lipschitz sensitivity of the map $\theta_{\text{wm}} \mapsto p_{(\theta_{\text{wm}},\theta_0)}(\cdot \mid x)$. In the worst case this constant can scale with a network Lipschitz factor (e.g., products of layer operator norms), which may grow exponentially in depth. In practice this is milder: restricting watermarking to later layers reduces the effective sensitivity, and empirical results from [13] demonstrate that the scaling can be much more moderate. Therefore, it is natural to make an assumption that there exists a model-dependent constant $C(p_\theta) > 0$ such that

$$\mathcal{I}_{jj}(\theta_{\text{wm}}) \le C(p_\theta).$$

Then we have $\mathbb{E}[D(\xi_\sigma)] \le \frac{C(p_\theta)\sigma^2 d_r}{2}$, plugging it into (6) and taking the supremum over $x \in \mathcal{X}$ yields

$$\mathbb{E}_{\xi_\sigma}\left[\sup_{x\in\mathcal{X}}\left\|p_{\theta(\xi_\sigma)}(\cdot \mid x) - p_\theta(\cdot \mid x)\right\|_{\mathrm{TV}}\right] \lesssim \sigma\sqrt{d_r}.$$

□

B.2 Proof of Proposition 2

Proof. Under the null hypothesis $\mathbf{H_0}$, for any $y \in \mathcal{Y}$, the key and the generated text y are independent of each other, i.e., $(\xi_\sigma, y) \sim \mathcal{N}(0, \sigma^2\mathbb{I}_{d_r}) \otimes q$ for some $q \in \Delta(\mathcal{Y})$. Therefore, the level of the test is given by

$$\begin{aligned}
\Pr_{\mathbf{H_0}}\big(\psi(y;\xi_\sigma)=1\big) &= \mathbb{E}_{\xi_\sigma\sim\mathcal{N}(0,\sigma^2\mathbb{I}_{d_r}),\, y\sim q}\left[\mathbb{I}\left\{\frac{\langle \xi_\sigma, \nabla_{\theta_{\mathrm{wm}}}\log q_{\theta'}(y)\rangle}{\sigma\|\nabla_{\theta_{\mathrm{wm}}}\log q_{\theta'}(y)\|_2} \ge \tau_\alpha\right\}\right] \\
&= \mathbb{E}_{y\sim q}\left[\Pr_{\xi_\sigma}\left(\frac{\langle \xi_\sigma, \nabla_{\theta_{\mathrm{wm}}}\log q_{\theta'}(y)\rangle}{\sigma\|\nabla_{\theta_{\mathrm{wm}}}\log q_{\theta'}(y)\|_2} \ge \tau_\alpha\right)\right] \\
&= \mathbb{E}_{y\sim q}\left[\Pr_{\xi_\sigma}\big(\psi(y;\xi_\sigma)\ge\tau_\alpha\big)\right] \\
&= 1-\Phi(\tau_\alpha) \;=\; \alpha,
\end{aligned}$$

where the last line is based on

$$\psi(y;\xi_\sigma) = \frac{\langle \xi_\sigma, \nabla_{\theta_{\mathrm{wm}}}\log q_{\theta'}(y)\rangle}{\sigma\|\nabla_{\theta_{\mathrm{wm}}}\log q_{\theta'}(y)\|_2} \sim \mathcal{N}(0,1)$$

for any vector $\nabla_{\theta_{\mathrm{wm}}}\log q_{\theta'}(y)$. The last equality is derived by plugging in $\tau_\alpha = \Phi^{-1}(1-\alpha)$. □

C A Stylized Linear-Softmax Analysis of MARKTUNE

Model and Notation. Let $\theta = (\theta_{\mathrm{wm}}, \theta_0)$ with last-layer (watermarked) parameter θ_{wm} and remaining parameter θ_0. For prompts x and responses y, we consider a similar linear-softmax model as in [13]:

$$p_\theta(y \mid x) \;\propto\; \exp\big\{\langle \theta_{\mathrm{wm}},\, \phi_{\theta_0}(x,y)\rangle\big\}. \tag{7}$$

A weight-editing watermark like GAUSSMARK fixes a key ξ and freezes the watermarked weights at $\theta_{\mathrm{wm}} + \xi$. We write $\theta(\xi) := (\theta_{\mathrm{wm}} + \xi,\ \theta_0)$. Let $p^\star(\cdot \mid x)$ denote a high-quality target distribution realized by some (not necessarily unique) weights $(\theta^\star_{\mathrm{wm}}, \theta^\star_0)$ in the same class, i.e.,

$$p^\star(y \mid x) \;\propto\; \exp\big\{\langle \theta^\star_{\mathrm{wm}},\, \phi_{\theta^\star_0}(x,y)\rangle\big\}.$$

We analyze the population cross-entropy risk

$$\mathcal{L}(\theta_0) \;:=\; \mathbb{E}_x\,\mathbb{E}_{y\sim p^\star(\cdot|x)}\big[-\log p_{\theta(\xi)}(y \mid x)\big], \tag{8}$$

and the watermark reward used by GAUSSMARK-MARKTUNE (unnormalized and conditioned on x for analysis)

$$\mathcal{R}(\theta_0) \;:=\; \mathbb{E}_x\,\mathbb{E}_{y\sim p_{\theta(\xi)}(\cdot|x)}\big[\langle \xi,\ \nabla_{\theta_{\mathrm{wm}}}\log q_{\theta'}(y \mid x)\rangle\big], \tag{9}$$

where $q_{\theta'}$ is the fixed white-box reference used for detection (as in the main text). Note that $\nabla_{\theta_{\mathrm{wm}}}\log q_{\theta'}(y \mid x) = \phi_{\theta'_0}(x,y) - \mu'(x)$ in the linear-softmax case of q, with $\mu'(x) := \mathbb{E}_{y\sim q_{\theta'}(\cdot|x)}[\phi_{\theta'_0}(x,y)]$. The dependence of $\mathcal{R}$ on θ_0 is via its on-policy expectation under $p_{\theta(\xi)}$.

Goal. We show that, for small $\kappa > 0$, there exists a perturbation of features $\delta\phi$ (realizable by moving θ_0) such that the adjusted model GAUSSMARK-MARKTUNE

$$p_{(\theta_{\mathrm{wm}}+\xi,\ \theta_0^{\mathrm{MT}})}(\cdot \mid x), \qquad \phi_{\theta_0^{\mathrm{MT}}}(x,y) \;:=\; \phi_{\theta_0^\star}(x,y) \;+\; \delta\phi(x,y),$$

(i) approaches $p^\star(\cdot \mid x)$ in cross-entropy with $\mathcal{L}(\theta_0^{\mathrm{MT}}) - \mathcal{L}(\theta_0^\star) = \mathcal{O}(\kappa^2)$, while (ii) achieves a *first-order* increase in the watermark reward $\mathcal{R}$ (hence a constant-variance mean shift in the test statistic), guaranteeing separability from the unwatermarked baseline.

Note that we will use the abbreviation $\delta\phi_x := \delta\phi(x,\cdot)$, so inner products such as $\langle b_x, \delta\phi_x\rangle$ denote $\mathbb{E}_{y\sim p^\star(\cdot|x)}[\,b(x,y)^\top\delta\phi(x,y)\,]$; that is, the y-dependence is absorbed into the Hilbert space notation.

Assumptions.

(A1) *Realizability & smoothness.* The target $p^\star$ is realized by $(\theta^\star_{\mathrm{wm}}, \theta^\star_0)$ and the feature map ϕ_{θ_0} is Fréchet-differentiable in θ_0. The parameterization is sufficiently expressive to realize the small feature perturbations $\delta\phi$ constructed below.

(A2) *Local quadratic expansion of $\mathcal{L}$.* Around $\theta^\star_0$, the cross-entropy admits the second-order expansion $\mathcal{L}(\theta^\star_0 + \delta) = \mathcal{L}(\theta^\star_0) + \frac{1}{2}\,\mathbb{E}_x\big[\langle \delta\phi_x,\ \Sigma_x\,\delta\phi_x\rangle\big] + o(\|\delta\phi\|^2)$, where $\delta\phi_x(\cdot) := \delta\phi(x,\cdot)$ is centered under $p^\star(\cdot \mid x)$, and $\Sigma_x := \mathrm{Cov}_{y\sim p^\star(\cdot|x)}\big[\phi_{\theta^\star_0}(x,y)\big]$ is a positive-definite covariance operator (the Fisher operator) associated with $p^\star$.

(A3) *Local first-order expansion of $\mathcal{R}$.* For small distributional perturbations induced by $\delta\phi$, the reward changes as $\mathcal{R}(\theta^\star_0 + \delta) = \mathcal{R}(\theta^\star_0) + \mathbb{E}_x\big[\langle b_x,\ \delta\phi_x\rangle\big] + o(\|\delta\phi\|)$, for some b_x as derived in Lemma 2.

(A4) *Centering constraint.* We restrict admissible $\delta\phi$ to satisfy $\mathbb{E}_{y\sim p^\star(\cdot|x)}\delta\phi(x,y) = 0$ for each x, so logits are perturbed only in identifiable directions.

Dual Optimization Problem. MARKTUNE balances reward and cross-entropy via a dual objective. Under (A2)–(A4), the local problem over centered $\delta\phi$ becomes

$$\max_{\delta\phi \text{ centered}}\ \mathbb{E}_x\big[\langle b_x,\ \delta\phi_x\rangle\big] \ -\ \frac{\lambda}{2}\,\mathbb{E}_x\big[\langle \delta\phi_x,\ \Sigma_x\,\delta\phi_x\rangle\big]. \tag{10}$$

This is a strictly concave quadratic program (in function space) with unique optimizer.

Lemma 1 (Closed-form optimizer)**.** *For each x, the unique maximizer of (10) is*

$$\delta\phi^\dagger_x \ =\ \frac{1}{\lambda}\,\Sigma_x^{-1}\,b_x,$$

and the optimal objective value equals $\frac{1}{2\lambda}\,\mathbb{E}_x\big[\langle b_x,\ \Sigma_x^{-1}\,b_x\rangle\big]$.

Proof. Fix x. The objective w.r.t. $\delta\phi_x$ is $J_x(\delta\phi_x) = \langle b_x, \delta\phi_x\rangle - \frac{\lambda}{2}\langle \delta\phi_x, \Sigma_x\delta\phi_x\rangle$. Differentiating in the (Hilbert) inner product and setting the first-order condition to zero gives $b_x - \lambda\,\Sigma_x\,\delta\phi_x = 0 \Rightarrow \delta\phi^\dagger_x = \frac{1}{\lambda}\,\Sigma_x^{-1} b_x$. Strict concavity follows from positive-definiteness of Σ_x. Substituting $\delta\phi^\dagger_x$ back yields $J_x(\delta\phi^\dagger_x) = \frac{1}{2\lambda}\langle b_x, \Sigma_x^{-1} b_x\rangle$. Averaging over x proves the claim. □

Lemma 2 (Reward gradient)**.** *Let $\mathcal{R}(\theta_0) = \mathbb{E}_x\,\mathbb{E}_{y\sim p_{\theta(\xi)}(\cdot|x)}\big[\langle \xi, \phi_{\theta'_0}(x,y) - \mu'(x)\rangle\big]$, where $q_{\theta'}$ (and thus θ'_0) is fixed. Consider a path $\theta_0(t) = \theta^\star_0 + t\,\delta\theta_0$ inducing features $\phi_{\theta_0(t)}$ and distributions $p_t(\cdot \mid x) := p_{(\theta_{\mathrm{wm}}+\xi,\ \theta_0(t))}(\cdot \mid x)$. Let $\delta\phi(x,y) := \frac{d}{dt}\phi_{\theta_0(t)}(x,y)\big|_{t=0}$ and impose the centering constraint $\mathbb{E}_{y\sim p^\star(\cdot|x)}\delta\phi(x,y) = 0$ for each x. Then the Gâteaux derivative of $\mathcal{R}$ at $t = 0$ is*

$$\frac{d}{dt}\Big|_{t=0}\mathcal{R}(\theta_0(t)) \ =\ \mathbb{E}_x\,\big\langle b_x,\ \delta\phi_x\big\rangle, \quad b_x(y) \ =\ (\theta_{\mathrm{wm}} + \xi)\,\Big(h(x,y) - \mathbb{E}_{y\sim p^\star(\cdot|x)}h(x,y)\Big),$$

where $h(x,y) := \langle \xi, \phi_{\theta'_0}(x,y) - \mu'(x)\rangle$, and $\langle\cdot,\cdot\rangle$ denotes the Hilbert-space inner product $\langle f_x, g_x\rangle := \mathbb{E}_{y\sim p^\star(\cdot|x)}[f(x,y)^\top g(x,y)]$.

Proof. By definition, $\mathcal{R}(\theta_0) = \mathbb{E}_x\,\mathbb{E}_{y\sim p_{\theta(\xi)}(\cdot|x)}[h(x,y)]$, with h independent of θ_0 because $q_{\theta'}$ is fixed. Along the path $t \mapsto \theta_0(t)$, for each fixed x the on-policy derivative is

$$\frac{d}{dt}\Big|_{t=0}\mathbb{E}_{y\sim p_t(\cdot|x)}[h(x,y)] \ =\ \mathbb{E}_{y\sim p_0(\cdot|x)}\big[h(x,y)\,s(x,y)\big],$$

where $s(x,y) := \frac{d}{dt}\log p_t(y \mid x)\big|_{t=0}$. In the linear-softmax model with frozen readout $\theta_{\mathrm{wm}} + \xi$ we have

$$\log p_t(y \mid x) = \langle \theta_{\mathrm{wm}} + \xi,\ \phi_{\theta_0(t)}(x,y)\rangle - \log\sum_{\tilde y}\exp\big\{\langle \theta_{\mathrm{wm}} + \xi,\ \phi_{\theta_0(t)}(x,\tilde y)\rangle\big\},$$

hence

$$s(x,y) = \big\langle \theta_{\mathrm{wm}} + \xi,\ \delta\phi(x,y)\big\rangle - \mathbb{E}_{\tilde y\sim p_0(\cdot|x)}\big\langle \theta_{\mathrm{wm}} + \xi,\ \delta\phi(x,\tilde y)\big\rangle.$$

Evaluating at $p_0 = p^\star$ and using the centering constraint gives $s(x,y) = \langle \theta_{\mathrm{wm}} + \xi,\ \delta\phi(x,y)\rangle$. Therefore

$$\frac{d}{dt}\Big|_{t=0} \mathcal{R}(\theta_0(t)) = \mathbb{E}_x\, \mathbb{E}_{y\sim p^\star(\cdot|x)}\big[h(x,y)\,\langle \theta_{\mathrm{wm}} + \xi,\ \delta\phi(x,y)\rangle\big].$$

Viewing $\delta\phi_x$ as the vector-valued function $y \mapsto \delta\phi(x,y)$ and recalling the inner product definition, this equals

$$\mathbb{E}_x \left\langle (\theta_{\mathrm{wm}} + \xi) \left(h(x,\cdot) - \mathbb{E}_{y\sim p^\star(\cdot|x)} h(x,y) \right),\ \delta\phi_x \right\rangle,$$

which proves the claim. □

Proposition 3 (Second-order CE cost and first-order reward gain)**.** *Let $\delta\phi_x^\dagger = \lambda^{-1}\Sigma_x^{-1} b_x$ be the optimizer of the local problem*

$$\max_{\delta\phi \text{ centered}} \ \mathbb{E}_x\big[\langle b_x,\ \delta\phi_x\rangle\big] - \frac{\lambda}{2}\,\mathbb{E}_x\big[\langle \delta\phi_x,\ \Sigma_x\,\delta\phi_x\rangle\big],$$

with b_x as in Lemma 2. Assume θ_0^{MT} realizes $\delta\phi^\dagger$. Let $\kappa := \lambda^{-1}$. Then, as $\kappa \to 0$,

$$\mathcal{L}(\theta_0^{\mathrm{MT}}) - \mathcal{L}(\theta_0^\star) \ = \ \frac{\kappa^2}{2}\,\mathbb{E}_x\big[\langle b_x,\ \Sigma_x^{-1}\, b_x\rangle\big] \ + \ o(\kappa^2),$$

and

$$\mathcal{R}(\theta_0^{\mathrm{MT}}) - \mathcal{R}(\theta_0^\star) \ = \ \kappa\,\mathbb{E}_x\big[\langle b_x,\ \Sigma_x^{-1}\, b_x\rangle\big] \ + \ o(\kappa).$$

Consequently the (normalized) detection statistic exhibits a mean shift linear in κ at essentially unchanged variance (maintaining level-α calibration), while the cross-entropy deviation from $p^\star$ grows only quadratically in κ.

Proof. By the quadratic CE expansion (A2),

$$\mathcal{L}(\theta_0^\star + \delta) - \mathcal{L}(\theta_0^\star) = \frac{1}{2}\,\mathbb{E}_x\big[\langle \delta\phi_x,\ \Sigma_x\,\delta\phi_x\rangle\big] + o(\|\delta\phi\|^2).$$

Plugging $\delta\phi_x^\dagger = \lambda^{-1}\Sigma_x^{-1} b_x$ yields

$$\frac{1}{2}\,\mathbb{E}_x\big[\langle \lambda^{-1}\Sigma_x^{-1} b_x,\ \Sigma_x\,\lambda^{-1}\Sigma_x^{-1} b_x\rangle\big] = \frac{1}{2\lambda^2}\,\mathbb{E}_x\big[\langle b_x,\ \Sigma_x^{-1}\, b_x\rangle\big] = \frac{\kappa^2}{2}\,\mathbb{E}_x\big[\langle b_x,\ \Sigma_x^{-1}\, b_x\rangle\big],$$

plus $o(\kappa^2)$. For the reward, the first-order expansion from Lemma 2 (assumption (A3) made explicit) gives

$$\mathcal{R}(\theta_0^\star+\delta) - \mathcal{R}(\theta_0^\star) = \mathbb{E}_x\big[\langle b_x,\ \delta\phi_x\rangle\big] + o(\|\delta\phi\|) = \lambda^{-1}\,\mathbb{E}_x\big[\langle b_x,\ \Sigma_x^{-1}\, b_x\rangle\big] + o(\lambda^{-1}) = \kappa\,\mathbb{E}_x\big[\langle b_x,\ \Sigma_x^{-1}\, b_x\rangle\big] + o(\kappa).$$

This proves the claim. □

Implications for Detectability and Separation from the Unwatermarked Model. Let $\psi(y;\xi)$ denote the (normalized) GAUSSMARK test statistic computed with $q_{\theta'}$. Under $\mathbf{H_0}$ (unwatermarked text), $\mathbb{E}[\psi] = 0$. Under $p_{(\theta_{\mathrm{wm}}+\xi,\ \theta_0^{\mathrm{MT}})}$, Proposition 3 implies a mean shift of order $\kappa\, C$ with $C = \mathbb{E}_x\big[\langle b_x,\ \Sigma_x^{-1}\, b_x\rangle\big] > 0$, at essentially unchanged variance (level-α test is preserved by the detection procedure). Therefore, for any fixed false-positive rate, the true-positive rate increases at *first order* in κ, while the cross-entropy distance to $p^\star$ grows only at *second order*. Since the final watermarked weights remains $\theta_{\mathrm{wm}} + \xi \neq \theta_{\mathrm{wm}}$, the fine-tuned model is statistically separated from the original unwatermarked $(\theta_{\mathrm{wm}}, \theta_0)$ in the watermark-sensitive direction, i.e., it stays "far" from the unwatermarked baseline while being "near" $p^\star$.

Remarks on Realizability. The derivation requires that the feature shift $\delta\phi^\dagger$ be realizable by a small change in θ_0. In over-parameterized transformers, local expressivity (wide, flat optimization basins) often suffices for such realizability. More generally, one may view $\delta\phi^\dagger$ as the Riesz representer of the linear functional b_x under the Fisher metric Σ_x; any parameterization that is locally surjective onto the centered feature subspace realizes $\delta\phi^\dagger$.

D Implementation Details

For the implementation of vanilla GAUSSMARK and GAUSSMARK-MARKTUNE, we select θ_{wm} to be the up-projection matrix of MLP in the layer 8. For the GRPO-style policy optimization Algorithm 2, we set the learning rate of the policy model as 1e-5 and the CE coefficient in MARKTUNE objective 2 to be $\lambda = 0.5$. We fine-tune the model for a total steps of 100 (3 GRPO iters per step) with a sampling batch size of 16 and temperature of 1.0, SFT batch size of 12 and group size of 4. AdamW optimizer is adopted with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight_decay $= 0.1$. We run the experiments on a 40GB NVIDIA A100 GPU.

E Supplemental Results

E.1 Empirical Results on Detectability

Figure 3 shows consistent trends as Figure 2a: GAUSSMARK-MARKTUNE achieves higher AUC and TPR across all FPR thresholds. Overall, these results demonstrate that MARKTUNE alleviates the distortion introduced by larger σ without compromising the watermark signal, and is particularly effective in short-sequence settings.

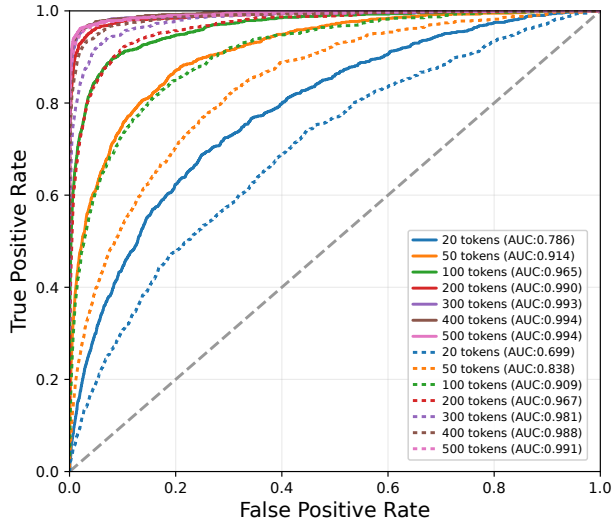
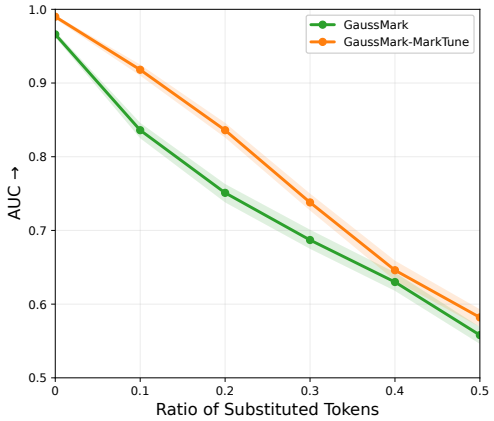


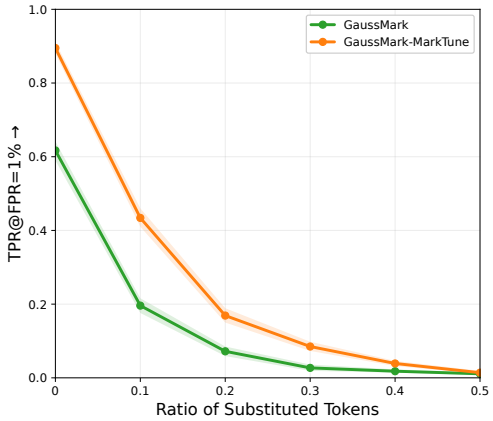
Figure 3: Watermark detectability under minimal distortion: ROC curves and the corresponding area under the curve (AUC) at different generated text lengths. The solid line denotes GAUSSMARK-MARKTUNE, while the dashed line denotes vanilla GAUSSMARK.

E.2 Empirical Results on Robustness

To comprehensively evaluate the robustness of our approach, we further consider a token-level attack (random token substitution) and a more challenging semantic-level attack (round-trip translation), which have been widely studied in prior work [16, 19, 23, 39, 40]. For the token-level attack, we randomly substitute a specific portion of tokens. Figure 4a and Figure 4b show that the detection performance decreases as the ratio of substituted tokens increases. However, We argue that such token-level perturbations are relatively crude and fall short of realistic threat models, since they noticeably degrade text quality. Therefore, following [16], we consider a more realistic yet challenging semantic-level attack. Specifically, we employ `Helsinki-NLP/opus-mt-tc-big-en-fr` and `Helsinki-NLP/opus-mt-tc-big-fr-en` [41] to translate watermarked text into French and back into English. Figure 5b presents the ROC curves and corresponding AUC values for each approach. Despite the higher difficulty of this attack, both methods maintain nontrivial detection power, demonstrating robustness under more realistic corruption scenarios. In particular, MARKTUNE improves the TPR across all FPR thresholds, demonstrating its strong potential to enhance the robustness

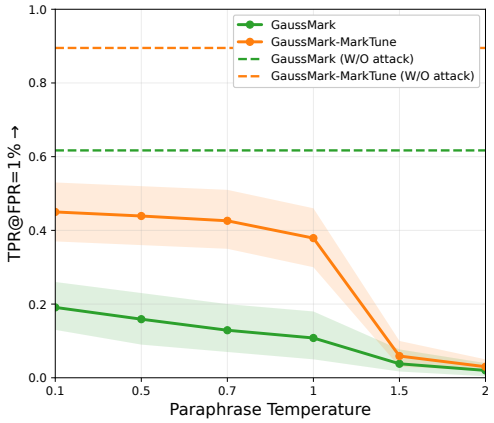


(a) AUC as a function of substitution ratio.

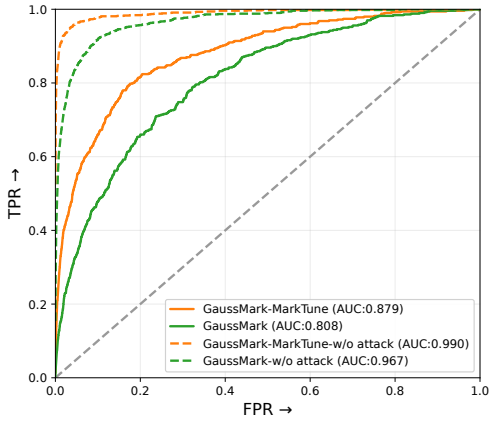


(b) TPR@FPR=1% as a function of substitution ratio.

Figure 4: Detection performance under random token substitution attack. All metrics are reported at an initial generation length of 200 tokens.



(a) TPR@FPR=1% as a function of paraphrasing temperature.



(b) ROC curves and corresponding AUC under roundtrip translation attack.

Figure 5: Detection performance under T5 model paraphrasing attack and roundtrip translation attack. All metrics are reported at an initial generation length of 200 tokens.

of weight-editing watermarking against roundtrip translation. For T5 model paraphrasing attack, Figure 5a also exhibits similar trend as Figure 2b.

E.3 Ablation Study

Table 2 compares SFT with our proposed on-policy fine-tuning framework MARKTUNE and evaluates the contribution of five components of GAUSSMARK-MARKTUNE: freezing θ_{wm}, the choice of θ_{wm} module, the choice of θ_{wm} layer, the key standard deviation σ, and the CE coefficient λ. The first three metrics are reported at sequence length 200. In the θ_{wm} module column, Up Proj. and Down Proj. refer to the up- and down-projection matrices of the MLP, each with 2.35M parameters. Atten. denotes the QKV projection matrices, with 1.77M parameters.

First of all, replacing on-policy, reward-conditioned policy updates (MARKTUNE) with plain SFT sharply weakens detectability (TPR from 0.895 to 0.627 at 1% FPR), even though text quality improves slightly. This suggests that the unfrozen weights undergo co-adaptation dynamics in the absence of watermark awareness during SFT. Additionally, freezing θ_{wm} yields a modest yet consistent improvement in both detectability and text quality, leading to more robust OOD generalization.

Table 2: Ablation study on components of GAUSSMARK-MARKTUNE. Each color block includes different choices of components. Best results are highlighted in **bold**.

MARKTUNE/SFT	Freeze θ_{wm}	θ_{wm} Module	θ_{wm} Layer	σ	λ	TPR@FPR=1%↑	AUC↑	PPL↓	Val. Loss↓	LAM.ACC↑	LAM.PPL↓
MARKTUNE	Yes	Up Proj.	8	0.1	0.5	**0.895**	**0.990**	13.59	3.118	0.306	37.97
SFT	Yes	Up Proj.	8	0.1	0.5	0.627	0.966	**12.03**	**3.088**	**0.310**	**37.21**
MARKTUNE	Yes	Up Proj.	8	0.1	0.5	**0.895**	**0.990**	**13.59**	**3.118**	**0.306**	**37.97**
MARKTUNE	No	Up Proj.	8	0.1	0.5	0.858	0.989	13.69	3.122	0.304	37.97
MARKTUNE	Yes	Up Proj.	8	0.1	0.5	**0.895**	**0.990**	13.59	3.118	0.306	37.97
MARKTUNE	Yes	Down Proj.	8	0.1	0.5	0.744	0.970	**13.49**	**3.115**	**0.309**	**37.64**
MARKTUNE	Yes	Atten.	8	0.1	0.5	0.107	0.807	14.69	3.149	0.286	49.58
MARKTUNE	Yes	Up Proj.	8	0.1	0.5	**0.895**	**0.990**	13.59	3.118	0.306	37.97
MARKTUNE	Yes	Up Proj.	10	0.1	0.5	0.837	0.982	**13.46**	**3.117**	**0.313**	**37.42**
MARKTUNE	Yes	Up Proj.	6	0.1	0.5	0.735	0.968	14.43	3.129	0.304	40.09
MARKTUNE	Yes	Up Proj.	8	0.1	0.5	0.895	0.990	13.59	3.118	0.306	37.97
MARKTUNE	Yes	Up Proj.	8	0.05	0.5	0.748	0.971	**12.83**	**3.091**	**0.327**	**36.29**
MARKTUNE	Yes	Up Proj.	8	0.2	0.5	**0.957**	**0.994**	14.89	3.257	0.289	47.40
MARKTUNE	Yes	Up Proj.	8	0.1	0.5	0.895	0.990	13.59	3.118	0.306	37.97
MARKTUNE	Yes	Up Proj.	8	0.1	0.2	**0.935**	**0.993**	15.03	3.152	0.296	46.15
MARKTUNE	Yes	Up Proj.	8	0.1	1.0	0.796	0.979	**12.15**	**3.101**	**0.316**	**36.67**

We also compare different module and layer choices for watermark embedding. Perturbations in the projection matrices of the MLP at deeper layers generally yield stronger quality-detectability trade-offs, with the up projection matrix at layer 8 showing the most favorable balance in our experiments. In contrast, injecting noise into earlier layers or into attention modules tends to substantially diminish the watermark signal.

Finally, a grid search over the hyperparameters σ and λ reveals a clear quality-detectability trade-off. Larger values of σ and smaller values of λ improve TPR@FPR=1% and AUC, but at the cost of higher perplexity and reduced LAMBADA accuracy. Notably, the setting $\sigma = 0.1, \lambda = 0.5$ achieves near–Pareto-optimal performance across all metrics.

F Limitations and Future Work

Several directions remain open for future work. First, it is worth applying MARKTUNE to larger open-weight LMs [1, 2, 35] and conducting comprehensive evaluations. Second, introducing adaptive hyperparameter search algorithm for MARKTUNE would be a promising direction to reduce the computational load of hyperparameter selection. Third, designing semantic embedding-awareness watermark signal reward and integrate it into MARKTUNE would potentially further enhance the watermark robustness against paraphrasing attacks. We hope these directions inspire future research toward robust and practical watermarking for open-weight LMs.