MarkTune: Improving the Quality-Detectability Trade-off in Model-Embedded LLM Watermarking

Yizhou Zhao
Georgia Institute of Technology
University of Pennsylvania
yzhao845@gatech.edu

Zhiwei Steven Wu
Carnegie Mellon University
zstevenwu@cmu.edu

Adam Block
Columbia University
adam.block@columbia.edu

Abstract

Language model watermarking schemes fall into two broad categories: *inference-time* methods, which modify the decoding process at generation time, and *model-embedded* methods, which encode a secret signal directly into the model weights. Inference-time watermarking methods incur additional inference overhead and are not applicable in the increasingly prevalent open-weight model settings. In contrast, model-embedded watermarks introduce no additional inference latency—generation uses the standard sampling pipeline—and are particularly well-suited to open-weight settings. However, existing model-embedded approaches such as GAUSSMARK face a fundamental quality-detectability trade-off: achieving strong detection power typically requires weight perturbations that noticeably degrade generation quality. Our key insight is that regularizing toward a high-quality reference distribution decouples quality preservation from proximity to the base model, enabling a better quality-detectability trade-off. Building on this insight, we introduce MARKTUNE, a theoretically grounded on-policy fine-tuning framework that uses the GAUSSMARK detection statistic as a reward while preserving generation quality through cross-entropy on high-quality human text or KL divergence to a stronger oracle model. Empirically, MARKTUNE substantially improves the quality-detectability frontier of GAUSSMARK, approaching the detectability of strong inference-time schemes while preserving generation quality and downstream task performance. MARKTUNE also remains robust under the evaluated text-editing and open-weight attacks and generalizes across datasets, retaining substantial detection power on corpora unseen during training. Code is available at https://github.com/zhyzmath/MarkTune-LLM-Watermarking.

1 Introduction

Language model watermarking aims to determine whether a given text was generated by a language model, a capability that is increasingly critical for maintaining academic integrity [41, 49, 66], mitigating large-scale misinformation [37, 69, 78], and protecting intellectual property in generative AI systems [38, 52, 75]. Most prior work formalizes watermarking as a statistical hypothesis testing problem [4, 35, 46], assuming a joint distribution over text and a secret watermarking key. Under this formulation, the goal is to generate text conditioned on a key such that the resulting dependence is reliably detectable, while satisfying quality constraints on the generated text. These constraints are frequently imposed via strict information-theoretic notions of non-distortion [10, 33, 44], for example by requiring watermarked and unwatermarked text distributions to be close in total variation.

Language model watermarking schemes fall into two broad categories: Inference-time schemes are primarily designed for *API-based deployment*, where the service provider controls the decoding process and can achieve strong statistical detectability with minimal generation quality degradation [1,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

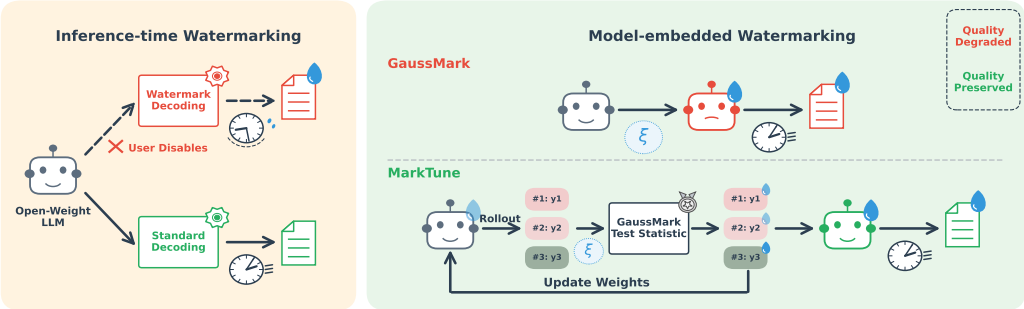


Figure 1: Overview of our framework compared to prior work [1, 4, 13, 42]. **Left**: Inference-time watermarking schemes break down on open-weight LMs because users can disable the decoding algorithm, and these methods often introduce substantial generation latency (indicates no latency; indicates extra latency). **Right**: Our approach, MARKTUNE, treats the GAUSSMARK test statistic as a reward and performs on-policy fine-tuning to embed a highly detectable yet quality-preserving watermark signal into the model's weights.

13, 42]. However, these methods introduce additional decoding overhead. Moreover, with the increasingly prevalent use of *open-weight* models [26, 28, 76], users can freely alter or bypass the decoding pipeline, making inference-time watermarking inapplicable. Model-embedded schemes avoid these limitations by embedding the watermark directly into the model weights, thereby enabling standard generation without additional inference latency and making them especially well suited to open-weight settings.

One recent model-embedded watermarking scheme that operates directly in weight space is GAUSSMARK [4], which adds keyed Gaussian noise to a subset of weight matrices, inducing a subtle but detectable shift in the output distribution. Unlike inference-time schemes based on hashed prefix keys, GAUSSMARK uses a single Gaussian weight-space key and full-sequence detection, avoiding the security–robustness trade-off [55]. Block et al. [4] showed that, with sufficiently small noise variance and carefully chosen parameters, GAUSSMARK achieves nontrivial detectability with negligible text-quality degradation. These results establish GAUSSMARK as a promising approach, but leave open how close it is to the optimal quality–detectability frontier and whether this trade-off can be improved through fine-grained, adaptive modifications to model weights. Prior work has also studied GAUSSMARK under continued fine-tuning attacks [24], motivating us to examine whether on-policy signal amplification preserves robustness when the released weights are subsequently updated.

We thus ask—*Can we design a watermarking scheme for open-weight LMs that preserves text quality and downstream task performance, achieves detection power comparable to inference-time watermarking, and remains robust to realistic text-editing and open-weight attacks?*

Our contribution. We answer the question above by proposing MARKTUNE, a theoretically principled on-policy fine-tuning framework that closes the quality-detectability gap between model-embedded and inference-time watermarking. Figure 1 contrasts our design with prior inference-time schemes. Our contributions are threefold:

- *Methodology.* MARKTUNE treats the detection statistic as an intrinsic watermark-signal reward and pairs it with a regularizer that anchors the trained model to an external high-quality distribution rather than to the base model, so that amplifying the watermark power and preserving generation quality are no longer antagonistic.
- *Theory.* We characterize a fundamental quality–detectability trade-off for GAUSSMARK: the perturbation strength and watermark subspace dimension jointly govern the quality cost and detection power. The MARKTUNE detector maintains exact Type-I error control. In a stylized linear-softmax regime, we further establish a detection-power bound for MARKTUNE and explain its improved trade-off through the lens of local optimization geometry.
- *Empirics.* On two open-weight LMs and two generation tasks, MARKTUNE consistently improves the quality-detectability trade-off of GAUSSMARK and approaches the detectability of strong

inference-time watermarks at comparable text quality. It preserves downstream-task performance, generalizes across datasets, and remains robust to realistic text-editing and open-weight attacks.

1.1 Related Work

Watermarking provides a proactive approach to identifying AI-generated content by embedding detectable signals while preserving content quality. Recent work spans text [4, 9, 10, 13, 27, 33, 42, 80], images [5, 11, 18, 19, 39, 62, 70, 77], video [20, 31, 32, 48, 79], audio [7, 47, 60, 61, 65], and tabular data [15, 29, 82–84]. Across these modalities, watermarking schemes must balance detectability and robustness with perceptual fidelity or downstream utility. Signals can be embedded through changes to model components, generative sampling, or the generated artifacts themselves, depending on the modality and deployment setting.

LM text watermarking methods fall into two broad categories: *inference-time* and *model-embedded* watermarking. Distortionary inference-time schemes modify the sampling process—for example, by biasing next-token selection toward a partitioned "green list" [42, 80]. While these methods offer statistical guarantees, they introduce noticeable text distortion and are vulnerable to realistic paraphrasing attacks [40, 59]. Nondistortionary inference-time schemes instead embed watermarks by controlling the pseudorandomness of sampling while preserving the original distribution. Examples include deterministic decoders based on the Gumbel-max trick [1] and inverse transform sampling [44], as well as unbiased variants of KGW using maximal coupling [73] or reweighting [71]. However, these approaches face challenges in large-scale deployment due to increased generation latency [4] and failures to preserve the joint distribution under key collisions [72]. Tournament-based watermarking mitigates latency overhead [13], but requires storing additional generated tokens, which limits its practicality.

Model-embedded watermarking directly encodes signals into model weights and can be further divided into training-based [27, 74] and weight-editing approaches [4, 9]. These methods are well suited to open-weight LMs and incur neither generation latency nor storage overhead. However, training-based schemes have limited generalization and lack rigorous statistical guarantees [27, 74], while existing weight-editing methods either require architectural changes [9] or depend on costly hyperparameter searches with limited gains in the quality-detectability trade-off [4].

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 specialize to the autoregressive generation setting, where prompts and responses are both represented as finite sequences of tokens from some vocabulary $\mathcal{V}$, so that $\mathcal{X}, \mathcal{Y} \subseteq \mathcal{V}^*$. In this case, the model generates a response one token at a time by sampling $y_1 \sim p(\cdot \mid x)$, then $y_t \sim p(\cdot \mid x, y_1, \ldots, y_{t-1})$ and concatenating the output tokens to form a response. Since our primary focus is on transformer-based language models, we parameterize the model with weights $\theta \in \Theta \subset \mathbb{R}^d$ and denote the resulting model as p_θ.

Hypothesis testing. As in Block et al. [4], Huang et al. [35], Li et al. [46], we formalize the notion of watermarking as a statistical hypothesis testing problem. Let Ξ denote the key space, from which the secret watermarking key ξ is sampled. 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. There are three phases in a weight-editing watermarking scheme: embedding, generation and detection. Formally, let Ξ be endowed with an associated distribution ρ. Given a language model p_θ, we first sample a key $\xi \sim \rho$ and obtain a watermarked model $p_{\theta(\xi)}$ by applying a weight-editing algorithm $\mathcal{A} : \Theta \times \Xi \to \Theta$, so that $\theta(\xi) = \mathcal{A}(\theta, \xi)$. The watermarked

text is then generated by sampling $y \sim p_{\theta(\xi)}(\cdot \mid x)$ using a prompt x. Detection is formulated as a hypothesis test, where $\mathbf{H_0} = \{\rho \otimes q | q \in \Delta(\mathcal{Y})\}$ denotes the set of joint distributions under which the key ξ and the text y are independent, and $\mathbf{H_A}$ corresponds to the joint distribution induced by the watermarking generation process.[1] The fundamental statistical difficulty of distinguishing $\mathbf{H_A}$ from $\mathbf{H_0}$ is governed by the total variation (TV) distance between the induced distributions under the two hypotheses [35], whose square is upper bounded by the well-known KL divergence up to a constant by Pinsker's inequality [6].

GaussMark. Due to the theoretical analysis below, in this work we focus on improving GAUSSMARK [4], a recently proposed weight-editing watermarking scheme. Given a language model $p_\theta : \mathcal{X} \to \Delta(\mathcal{Y})$, GAUSSMARK partitions the weights 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 fixed reference model, denoted as $p_{\theta_{\text{ref}}}$ to avoid confusion, 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 unchanged. To detect the watermark, GAUSSMARK uses the following test statistic:

$$\psi(y, \xi_\sigma \mid x) = \frac{\langle \xi_\sigma, \nabla_{\theta_{\text{wm}}} \log p_{\theta_{\text{ref}}}(y \mid x) \rangle}{\sigma \|\nabla_{\theta_{\text{wm}}} \log p_{\theta_{\text{ref}}}(y \mid x)\|_2}. \tag{1}$$

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 MARKTUNE: On-Policy Fine-Tuning with Watermark Signal Rewards

In this section, we introduce our proposed watermarking scheme, MARKTUNE, a theoretically principled approach that enhances the quality-detectability trade-off of GAUSSMARK through on-policy fine-tuning with watermark signal rewards. To motivate MARKTUNE, we first present the following result on the distortion introduced by GAUSSMARK.

Proposition 3.1. *Let p_θ be a language model with parameters $\theta = (\theta_{\text{wm}}, \theta_0) \in \Theta$, where $\theta_{\text{wm}} \in \mathbb{R}^{d_r}$ is the subset of parameters to which the* GAUSSMARK *is applied. Assume that the map $\theta' \mapsto \mathrm{D_{KL}}\big(p_{\theta'}(\cdot \mid x) \,\big\|\, p_\theta(\cdot \mid x)\big)$ is twice continuously differentiable with uniformly bounded Hessian.*[2] *If $\xi_\sigma \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_r})$ and $\theta(\xi_\sigma) = (\theta_{\text{wm}} + \xi_\sigma, \theta_0)$, then for any prompt x there is a constant C depending on the Hessian bound such that*

$$\mathbb{E}_{\xi_\sigma}\left[\mathrm{D_{KL}}\Big(p_{\theta(\xi_\sigma)}(\cdot \mid x) \,\Big\|\, p_\theta(\cdot \mid x)\Big)\right] \le C\, \sigma^2 d_r. \tag{2}$$

Proposition 3.1 characterizes the quality side of GAUSSMARK: the expected KL divergence between the perturbed and base models is bounded by $C\sigma^2 d_r$. Due to space constraints, we defer the detectability side to Appendix B.3, where we show that the type-II error of the GAUSSMARK test decays exponentially in d_r at a rate governed by σ. Together, these two results expose a fundamental quality-detectability trade-off for GAUSSMARK: increasing σ or the watermark subspace dimension d_r amplifies the detection signal but inevitably inflates the upper bound on the resulting quality 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 the high-quality text but instead is a learned model with imperfections. Consequently, closeness *to the base model* is not a necessary condition for achieving high-quality generation; rather, one should prefer closeness to some optimal distribution $p_{\theta^\star}$ that better expresses the true distribution of observed text. Such intuition is expressed in Figure 2, where the base model p_θ is suboptimal compared to the high-quality reference model $p_{\theta^\star}$; while GAUSSMARK attempts to maintain generation quality by remaining close to θ, our proposed method MARKTUNE (outlined below) instead seeks parameters that remain close to $\theta^\star$.

[1] 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**.

[2] This condition can be relaxed to local continuity and bounding of the Hessian at the cost of conditioning on the high probability event that $\theta(\xi)$ remains in this neighborhood.

Second, KL divergence is a particularly stringent measure of distributional distortion: it upper-bounds worst-case deviations over all events, far exceeding what is necessary to capture human-perceived quality. Modern LMs are heavily over-parameterized, with loss landscapes that contain broad, flat basins [8]. Consequently, given a base model p_θ, there typically exist alternative weight configurations $p_{\theta^\star}$ that exhibit a significant TV distance from p_θ but that yield essentially the same generation quality on general or domain-specific tasks. This perspective is also depicted in Figure 2, where both GAUSSMARK and MARKTUNE achieve similar watermark strength (denoted by the horizontal distance from θ) but our proposed MARKTUNE maintains better generation quality by remaining in the basin of the high-quality reference model $\theta^\star$.

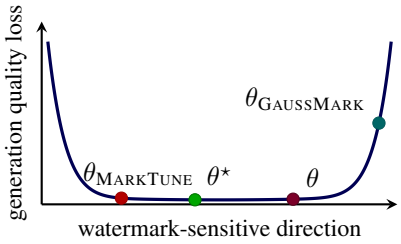


Figure 2: Stylized one-dimensional landscape along a watermark-sensitive direction. Both $\theta_{\text{GAUSSMARK}}$ and θ_{MARKTUNE} lie at nontrivial distances from the base model θ along this direction, leading to significant watermark detectability, but θ_{MARKTUNE} resides within the flat high-quality basin around $\theta^\star$ and therefore incurs substantially less quality degradation than $\theta_{\text{GAUSSMARK}}$.

Design motivation. Inspired by the success of Reinforcement Learning from Human Feedback (RLHF) [53] and Reinforcement Learning with Verifiable Rewards (RLVR) [28] in enhancing LLM capabilities, we propose MARKTUNE, an on-policy fine-tuning framework that uses the GAUSSMARK test statistic as its reward. Unlike the rewards in RLHF and RLVR, which are typically produced by trained reward models, domain-specific verifiers, or gold-standard solutions, this test statistic is an intrinsic quantity that can be easily computed using a watermark key. Following prior work, we include a regularization term in the optimization objective to preserve generation quality. A central design choice is *what to regularize toward*. Conventional RL objectives [28, 53, 63] anchor the trained model to the base model via KL divergence, but for watermarking this anchor is counterproductive: increasing the watermark signal inherently moves the trained weights away from the base model, so a base-model anchor pulls directly against the reward and this distribution-preservation constraint can limit improvements over the existing GAUSSMARK trade-off frontier. We instead anchor to an external high-quality reference distribution, instantiated either as the cross-entropy on human text or as the KL divergence to a stronger oracle from the same model family. Both instantiations decouple quality preservation from base-model proximity: they enlarge the search space for watermark-aware weight updates and bias the training dynamics toward watermark-sensitive directions along which quality decays only slowly, yielding a better quality-detectability frontier. We empirically validate this principle in Appendix E.3 (Table 9), where replacing the anchor with the standard base-model KL fails to improve over GAUSSMARK at matched quality.

Optimization with dual objectives. Given a base model p_θ with $\theta = (\theta_{\text{wm}}, \theta_0)$ and a watermark key $\xi_\sigma \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_r})$ for small σ, MARKTUNE first applies a GAUSSMARK soft activation, $\theta(\xi_\sigma) = (\theta_{\text{wm}} + \xi_\sigma, \theta_0)$, which induces a weak watermark signal while preserving generation quality. This procedure yields two distinct models: a fixed copy of the base model, denoted as $p_{\theta_{\text{ref}}}$, which is used to compute the reward (i.e., the test statistic), and an updated data generating model $p_{\theta(\xi_\sigma)}$ (with parameter $\theta(\xi_\sigma)$) that is optimized to adapt the watermark embedding. The training objective for $p_{\theta(\xi_\sigma)}$ is to maximize

$$\mathcal{J}(\theta(\xi_\sigma)) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim p_{\theta(\xi_\sigma)}(\cdot|x)}\left[\frac{\langle \xi_\sigma, \nabla_{\theta_{\text{wm}}} \log p_{\theta_{\text{ref}}}(y)\rangle}{\sigma \|\nabla_{\theta_{\text{wm}}} \log p_{\theta_{\text{ref}}}(y)\|_2}\right] - \lambda \mathcal{L}_{\text{reg}}(\theta(\xi_\sigma); x, y)\,. \tag{3}$$

where the first term is the watermark signal reward $\mathcal{R}_{\text{wm}}(y; \xi_\sigma)$, formulated using the GAUSSMARK test statistic.[3] The term $\mathcal{L}_{\text{reg}}$ is a regularization term, which can be either the cross-entropy loss $\mathcal{L}_{\text{ce}} = -\frac{1}{T}\sum_{t=1}^{T} \log p_{\theta(\xi_\sigma)}(x_t \mid x_{<t})$ or the KL divergence to an oracle model $\mathcal{L}_{\text{kl}} = \mathrm{D}_{\text{KL}}\big(p_{\theta(\xi_\sigma)}(\cdot \mid x) \,\|\, p_{\text{oracle}}(\cdot \mid x)\big)$.[4] $\lambda > 0$ denotes a hyperparameter that balances the watermark reward against the regularization on text quality degradation. By rewarding generated

[3] We ignore the prompt x to match the real-world detection.

[4] We present an ablation study on the choice of regularization term in Appendix E.2.

Algorithm 1 MARKTUNE with GAUSSMARK activation

1: **Input:** LM p_θ with $\theta = (\theta_{\text{wm}}, \theta_0)$, strength $\sigma > 0$, dataset $\mathcal{D}$, regularization coefficient $\lambda > 0$.
2: Sample a key $\xi_\sigma \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_r})$ and apply GAUSSMARK to obtain $\theta(\xi_\sigma) = (\theta_{\text{wm}} + \xi_\sigma, \theta_0)$.
3: Store p_θ as the fixed reference model $p_{\theta_{\text{ref}}}$.
4: Conduct GRPO Algorithm 2 for $\theta(\xi_\sigma)$ to obtain $\theta^\star(\xi_\sigma)$ via maximizing the objective (3).
5: **Output:** Watermarked weights $\theta^\star(\xi_\sigma)$.

samples $y \sim p_{\theta(\xi_\sigma)}(\cdot \mid x)$ that produce stronger watermark signals, while penalizing deviations from high-quality text through $\mathcal{L}_{\text{reg}}$, this dual-objective, on-policy fine-tuning framework improves the quality-detectability trade-off of GAUSSMARK. The high-level procedure of MARKTUNE is summarized in Algorithm 1. For practical implementation, we employ GRPO [64] to maximize the overall objective (3), with algorithm details provided in Appendix A. We further examine the roles of Gaussian initialization and quality regularization using direct-GRPO and quality-only fine-tuning controls, respectively (Appendices E.1 and E.3).

Remark 3.2. We emphasize that, in principle, MARKTUNE can serve as a general on-policy fine-tuning framework for improving the quality-detectability trade-off of a broad class of watermarking schemes, by replacing the first term in (3) with the corresponding test statistic or any other computable detection statistic as the reward. We choose GAUSSMARK in this work due to the theoretical motivation outlined below. An empirical demonstration of MARKTUNE instantiated with the KGW z-score as reward is provided in Appendix E.6 (Table 13). Extensions to semantic-level statistics that involve detection-time resampling are nontrivial and discussed in Appendix G.

Type I error control. For watermark detection, we inherit the procedure in GAUSSMARK. Since the generated text y remains independent of ξ_σ under $\mathbf{H_0}$, Proposition 3.3, whose proof is in Appendix B.2, provides rigorous statistical guarantees on the false positive rate.

Proposition 3.3. *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 $\mathbb{I}\left\{\frac{\langle \xi_\sigma, \nabla_{\theta_{\text{wm}}} \log p_{\theta_{\text{ref}}}(y)\rangle}{\sigma \|\nabla_{\theta_{\text{wm}}} \log p_{\theta_{\text{ref}}}(y)\|_2} \geq \tau_\alpha\right\}$ has level α.*

Informal analysis for MARKTUNE's improvement. We can gain intuition for why MARKTUNE improves upon GAUSSMARK by considering the linear-softmax regime, a common analytically tractable proxy for studying language models [4, 22, 34]. Suppose the "language model" takes the form $p_\theta(y \mid x) \propto \exp(\langle \theta, \Phi(x,y)\rangle)$, where $\Phi(x,y)$ is a fixed feature map. This setting can be viewed as an approximation in which all but the final layer of a transformer are frozen. Let $\theta^\star$ denote the parameter of the optimal language model and θ that of the base model. First, consider $\mathcal{L}_{\text{kl}} = \mathrm{D}_{\text{KL}}\big(p_{\theta(\xi_\sigma)} \,\|\, p_{\theta^\star}\big)$, under which the objective in (3) becomes a reverse KL-regularized policy optimization problem over $\theta(\xi_\sigma)$. Applying the standard reparameterization trick [57], if optimization is carried to completion (and ignoring the rescaling in (3) for simplicity), we obtain $p_{\theta_{\text{MARKTUNE}}}(y \mid x) \propto p_{\theta^\star}(y \mid x) \exp(\langle \xi, \Phi(x,y)\rangle)$. Under the linear-softmax assumption, this yields $p_{\theta_{\text{MARKTUNE}}}(y \mid x) \propto \exp(\langle \theta^\star + \xi, \Phi(x,y)\rangle)$, whereas GAUSSMARK satisfies $p_{\theta_{\text{GAUSSMARK}}}(y \mid x) \propto \exp(\langle \theta + \xi, \Phi(x,y)\rangle)$. Thus, at a high level, MARKTUNE effectively applies GAUSSMARK to the *optimal parameter* rather than the base-model parameter, providing intuition for its improved quality-detectability frontier. Appendix B.4 provides a conditional detection power bound for this idealized population optimizer.

The cross-entropy regularizer $\mathcal{L}_{\text{ce}}$ corresponds, up to an additive constant, to the forward KL $\mathrm{D}_{\text{KL}}\big(p_{\theta^\star} \,\|\, p_{\theta(\xi_\sigma)}\big)$, for which the reparameterization trick does not directly apply. We therefore analyze this case locally in Appendix C. In a neighborhood of the high-quality distribution $p_{\theta^\star}$, there exists a deviation $\delta\theta$ along a watermark-sensitive direction such that the watermark reward increases at *first order* in $\kappa = \frac{1}{\lambda}$, while the corresponding increase in cross-entropy (or KL) relative to $p_{\theta^\star}$ is only *second order*. This provides complementary theoretical insight into how MARKTUNE can adaptively embed a watermark while preserving model quality.

4 Experiments

In this section, we evaluate MARKTUNE on two language models and datasets to demonstrate its applicability across models and generalization across tasks. Our evaluation focuses on four aspects:

1) generation quality and watermark detectability compared to existing inference-time and model-embedded schemes; 2) downstream task performance; 3) robustness to text-editing attacks; and 4) robustness to fine-tuning attacks. **All experiments use *prompt-agnostic* detection**, reflecting realistic deployment scenarios where prompts are unavailable to the detector. We instantiate the MARKTUNE detector using the statistic in (1) without prompt access:

$$\psi(y, \xi_\sigma) = \frac{\langle \xi_\sigma, \nabla_{\theta_{\text{wm}}} \log p_{\theta_{\text{ref}}}(y) \rangle}{\sigma \|\nabla_{\theta_{\text{wm}}} \log p_{\theta_{\text{ref}}}(y)\|_2},$$

where gradients are computed with respect to the base model log-likelihood of the full sequence y.

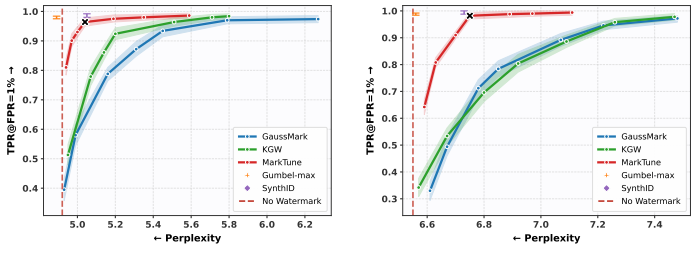
4.1 Experimental Setup

Models and datasets. We evaluate MARKTUNE on two open-weight LMs: Qwen3-4B-Base [76] and Llama2-7B [68]. Qwen3-4B-Base is a relatively new model, while Llama2-7B is widely used in prior watermarking studies [27, 44, 71, 74, 81]. We adopt the sampling temperatures recommended by each model's technical report: $T = 0.7$ for Qwen3-4B-Base [76] and $T = 1.0$ for Llama2-7B [68]. We evaluate on two generation tasks: text completion and question answering. For text completion, prompts are drawn from the **RealNewsLike** split of **C4** [58]; for long-form question answering, we use **ELI5** [14]. For each task, we sample 500 prompts and generate 200-token responses. MARKTUNE is fine-tuned on a subset of **OpenWebText** [25], using prompt lengths of 64-256 tokens for on-policy sampling. The training data are disjoint from and substantially different than the evaluation datasets, demonstrating strong cross-domain generalization.

Baselines and metrics. We benchmark MARKTUNE against representative watermarking schemes, including inference-time methods KGW [42] with two hyperparameter settings (KGW-1/2), OpenAI's Gumbel-max [1], Google's SynthID [13], and model-embedded approaches such as KGW-D [27] (logit distillation of KGW) and GAUSSMARK [4]. We exclude RL-based watermarking [74], as it failed to achieve nontrivial detection under multinomial sampling across datasets. For a fair comparison, we tune each method to attain nontrivial detectability with minimal quality degradation. Detectability is measured by the true positive rate at a fixed false positive rate of 1% (TPR@FPR=1%) and by the area under the ROC curve (AUC). Generation quality is assessed using perplexity (PPL), computed with Qwen3-8B and Llama2-13B as oracle language models for the corresponding evaluated models. We measure repetitiveness using Seq-rep-3, defined as the average rate of duplicate 3-grams within a sequence, and report MAUVE scores to quantify distributional similarity between model outputs and human text [56]. Implementation details are provided in Appendix D. Ultra-low-FPR detection results and an additional model-embedded baseline are reported in Appendices F.1 and F.3; sensitivity to watermarked parameters, decoding strategy, and output length is analyzed in Appendices E.7 and F.4.

4.2 Main Results

Table 1 reports the generation quality and detection performance of different watermarking methods across models and datasets, while Figures 3 and 4 respectively summarize the resulting quality-detectability trade-off and the wall-clock generation/detection times on **C4-RealNewsLike**.[5] For each method, we sweep a scheme-specific parameter to trace the trade-off curve, except for Gumbel-max and SynthID, which are non-distortionary and therefore yield a single TPR-PPL point. We observe that **MARKTUNE consistently outperforms other model-embedded watermarking approaches**



Qwen3-4B-Base Llama2-7B

Figure 3: Quality-detectability trade-off at 200 generated tokens. Black '×' marks the MARKTUNE configuration used elsewhere.

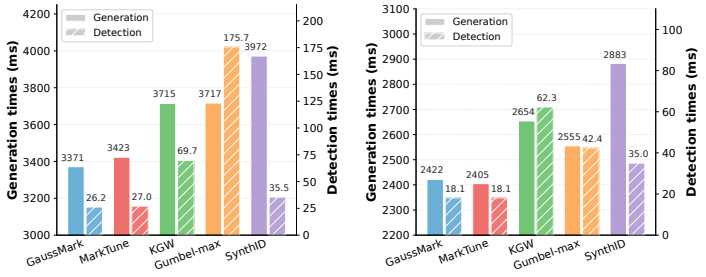
[5] We omit the trade-off curve of KGW-D since its performance is upper bounded by KGW.

Table 1: Generation quality and watermark detectability across methods, best-performing model-embedded watermarks are shown in **bold**.

Method	C4-RealNewsLike					ELI5				
	AUC↑	TPR↑	PPL↓	Seq-rep-3↓	MAUVE↑	AUC↑	TPR↑	PPL↓	Seq-rep-3↓	MAUVE↑
Qwen3-4B-Base, T=0.7										
No Watermark	0.500	0.010	$4.92_{1.34}$	0.04	0.98	0.500	0.010	$2.67_{0.94}$	0.03	0.94
KGW-1	0.996	0.964	$5.51_{1.56}$	0.04	0.98	0.971	0.710	$3.04_{1.12}$	0.04	0.92
KGW-2	1.000	0.998	$6.65_{1.91}$	0.04	0.96	0.996	0.988	$3.69_{1.51}$	0.03	0.87
Gumbel-max	0.995	0.980	$4.89_{2.07}$	0.18	0.74	0.991	0.938	$2.71_{1.01}$	0.13	0.77
SynthID	0.996	0.986	$5.05_{1.54}$	0.04	0.96	0.993	0.946	$2.76_{0.95}$	0.06	0.91
KGW-D	0.981	0.762	$5.54_{1.53}$	0.04	0.98	0.941	0.542	$3.09_{1.17}$	0.04	0.92
GaussMark	0.983	0.788	$5.16_{1.54}$	0.04	**0.98**	0.979	0.734	$2.83_{1.07}$	0.03	0.93
MarkTune	**0.995**	**0.964**	$\mathbf{5.04_{1.37}}$	**0.03**	0.97	**0.996**	**0.942**	$\mathbf{2.78_{1.03}}$	**0.03**	**0.93**
Llama2-7B, T=1.0										
No Watermark	0.500	0.010	$6.55_{1.80}$	0.01	0.98	0.500	0.010	$7.86_{2.23}$	0.02	0.97
KGW-1	0.991	0.886	$7.09_{2.04}$	0.02	0.96	0.993	0.946	$8.47_{2.39}$	0.04	0.94
KGW-2	0.997	0.996	$8.14_{2.36}$	0.01	0.95	1.000	1.000	$9.61_{2.68}$	0.03	0.91
Gumbel-max	0.997	0.988	$6.56_{1.84}$	0.14	0.86	0.998	0.994	$7.98_{2.29}$	0.18	0.81
SynthID	0.995	0.994	$6.73_{1.94}$	0.03	0.94	0.997	0.988	$8.06_{2.41}$	0.09	0.89
KGW-D	0.945	0.582	$7.04_{1.98}$	0.02	0.98	0.969	0.754	$8.53_{2.41}$	0.05	0.96
GaussMark	0.974	0.712	$6.78_{1.91}$	0.01	**0.98**	0.958	0.622	$\mathbf{8.13_{2.15}}$	0.02	0.94
MarkTune	**0.997**	**0.982**	$\mathbf{6.75_{1.96}}$	**0.01**	0.96	**0.993**	**0.930**	$8.19_{2.27}$	**0.01**	**0.96**

in detectability while introducing only minimal degradation in text quality (all three quality metrics remain close to the unwatermarked baseline), pushing the trade-off frontier strictly beyond GAUSSMARK and KGW and matching inference-time watermarking methods. KGW attains high detectability but at the cost of substantial perplexity degradation. In contrast, Gumbel-max generally offers strong detectability and low perplexity, but suffers from notable repetitiveness due to its deterministic decoding procedure. SynthID generally performs well in both detectability and generation quality.

On the computation time side (Figure 4), MARKTUNE matches GAUSSMARK at both generation and detection time and is substantially faster than every inference-time baseline on both axes, so its quality-detectability gains incur no inference-time overhead. Computation time benchmark details are deferred to Appendix D.4. Detector access requirements and measured peak GPU memory are reported in Appendix D.5.



Qwen3-4B-Base Llama2-7B

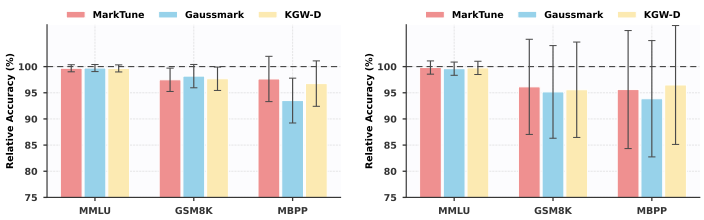
Figure 4: Mean wall-clock times on the first 100 **C4-RealNewsLike** prompts (200-token prefix, 200 generated tokens).

This comparison concerns generation and detection after watermark embedding. MARKTUNE additionally incurs a one-time offline GRPO training cost, reported separately in Appendix D.6.

Impact on downstream task performance. Because watermarking strategies alter the output distributions of LMs, it is essential to ensure that the utility of the underlying model is not compromised. Although the results in Table 1 indicate minimal degradation in generation quality—measured by perplexity, Seq-rep-3, and MAUVE—recent work [2, 4, 13] has shown that such metrics do not reliably predict the performance degradation introduced by watermarking on downstream tasks. Therefore, we further evaluate the watermarked models on downstream tasks using the Language Model Evaluation Harness [23] across three benchmarks: MMLU [30] for general knowledge and reasoning, GSM8K [12] for mathematical problem solving, and MBPP [3] for coding tasks.

Figure 5 reports the accuracy of three model-embedded watermarking schemes on these benchmarks, relative to the unwatermarked model.

On the MMLU benchmark, all methods exhibit negligible accuracy degradation, indicating that they preserve the model's general capabilities. On the GSM8K and MBPP benchmarks, MARKTUNE ranks first or second, substantially improving over GAUSSMARK, especially on MBPP. We attribute this to the fact that MARKTUNE more effectively guides the training dynamics toward "directions" that simultaneously enhance the watermark signal and preserve utility, starting from only a slight weight perturbation. In contrast, GAUSSMARK requires a much larger perturbation to produce a nontrivial watermark signal, which can negatively affect downstream performance. We provide watermarked examples in Appendix H for generation quality illustration beyond these quantitative metrics.



(a) Qwen3-4B-Base (b) Llama2-7B

Figure 5: Relative downstream task accuracy compared to unwatermarked models across general, math, and coding benchmarks.

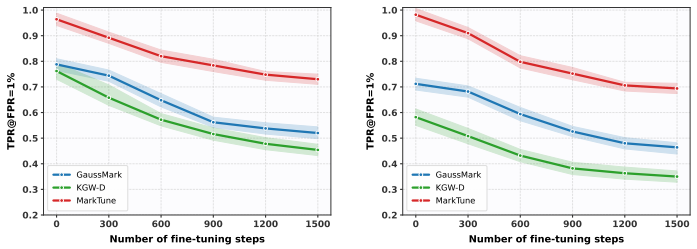
4.3 Robustness Evaluation

Robustness to text-editing attacks. Text editing is common in practical LM use and can weaken watermark signals. We evaluate robustness against four editing attacks: word deletion, synonym substitution, paraphrasing, and roundtrip translation. As shown in Table 2, for deletion and substitution, we randomly modify a fraction of watermarked tokens, sampling synonyms from WordNet [50]. Paraphrasing is performed using two Dipper configurations with varying lexical and order diversity [43], while a stronger paraphrasing attack uses French–English roundtrip translation with MarianMT-Model [67]. Under word deletion and synonym substitution, KGW-D often performs strongly among model-embedded methods, likely because it is distilled from the one-token-prefix KGW watermark. While MARKTUNE consistently improves over GAUSSMARK, its watermark signal still decays at higher modification rates. However, such token-level attacks are relatively crude and can substantially degrade text quality. To quantify this effect, Table 24 (deferred to the appendix for space) reports the perplexity of text after each editing attack. Token-level attacks substantially degrade text quality, whereas paraphrasing attacks better preserve quality while reducing the detectability of alternative watermarks. Notably, MARKTUNE remains robust under these paraphrasing attacks. Thus, focusing on the more realistic paraphrasing setting, MARKTUNE outperforms all watermarking baselines, including strong inference-time approaches, demonstrating its robustness and practical applicability under adversarial text editing.

Robustness to open-weight attacks. Given the open-weight deployment setting, we also evaluate the robustness of weight-embedded watermarks to fine-tuning attacks. With the growing adoption of LoRA as a PEFT technique, we consider an adversary who attempts to remove the watermark by further fine-tuning the model with LoRA on a subset of **OpenWebText** that is disjoint from the dataset used during MARKTUNE training, following the setup in prior work [24].

As shown in Figure 6, although MARKTUNE enhances the watermark via fine-grained on-policy weight updates, it does not exhibit a faster TPR decay rate than the other two methods. Owing to its strong initial watermark signal, it retains substantial detection power even after 1500 fine-tuning steps, in contrast to GAUSSMARK and KGW-D.

Beyond LoRA fine-tuning, we evaluate whether common



(a) Qwen3-4B-Base (b) Llama2-7B

Figure 6: Detectability (TPR@1%FPR) decay under LoRA fine-tuning attack.

Table 2: Watermark detectability under text-editing and paraphrasing attacks for different methods on **C4-RealNewsLike**. "Del-n" and "Sub-n" denote attacks that delete or substitute a proportion of n tokens, respectively. "Dipper-1" refers to Dipper paraphrasing with lexical and order diversity set to 20, while "Dipper-2" uses a diversity level of 60. "Translate" denotes the roundtrip translation attack described in Section 4.3. Best performances among model-embedded watermarks are shown in **bold**.

Method	Del-0.2		Del-0.5		Sub-0.2		Sub-0.5		Dipper-1		Dipper-2		Translate	
	AUC↑	TPR↑	AUC↑	TPR↑	AUC↑	TPR↑	AUC↑	TPR↑	AUC↑	TPR↑	AUC↑	TPR↑	AUC↑	TPR↑
Qwen3-4B-Base, T=0.7														
KGW-1	0.881	0.330	0.617	0.040	0.873	0.324	0.638	0.048	0.771	0.142	0.573	0.026	0.831	0.242
KGW-2	0.975	0.746	0.697	0.082	0.986	0.826	0.747	0.106	0.884	0.372	0.616	0.038	0.928	0.552
Gumbel-max	0.993	0.948	0.778	0.162	0.993	0.922	0.840	0.348	0.945	0.748	0.721	0.146	0.955	0.846
SynthID	0.987	0.890	0.780	0.094	0.986	0.846	0.834	0.182	0.941	0.576	0.700	0.074	0.968	0.818
KGW-D	**0.947**	**0.582**	**0.851**	**0.196**	0.939	**0.578**	**0.837**	**0.190**	0.855	0.274	0.589	0.030	0.912	0.404
GaussMark	0.908	0.366	0.622	0.048	0.919	0.384	0.775	0.100	0.915	0.438	0.795	0.138	0.948	0.574
MarkTune	0.921	0.484	0.744	0.068	**0.948**	0.552	0.783	0.110	**0.977**	**0.852**	**0.851**	**0.254**	**0.973**	**0.858**
Llama2-7B, T=1.0														
KGW-1	0.853	0.232	0.581	0.024	0.878	0.300	0.638	0.030	0.757	0.116	0.551	0.014	0.777	0.146
KGW-2	0.975	0.712	0.711	0.060	0.980	0.756	0.772	0.126	0.862	0.322	0.614	0.032	0.922	0.448
Gumbel-max	0.993	0.924	0.790	0.224	0.993	0.950	0.862	0.296	0.950	0.700	0.711	0.148	0.961	0.796
SynthID	0.979	0.774	0.777	0.114	0.987	0.862	0.843	0.140	0.943	0.552	0.710	0.084	0.949	0.612
KGW-D	0.914	0.410	**0.803**	**0.148**	0.887	0.406	0.759	**0.144**	0.814	0.238	0.590	0.026	0.889	0.324
GaussMark	0.835	0.242	0.617	0.036	0.853	0.228	0.759	0.098	0.857	0.240	0.687	0.060	0.925	0.436
MarkTune	**0.928**	**0.506**	0.783	0.080	**0.944**	**0.578**	**0.809**	0.122	**0.961**	**0.836**	**0.859**	**0.270**	**0.966**	**0.834**

modifications to released weights weaken the watermark while preserving model utility. Under the evaluated INT8 quantization, WANDA pruning, and SLERP merging settings, MARKTUNE retains nontrivial detectability with generation quality close to the unmodified watermarked model (Appendix F.6). We also evaluate logit-based distillation from a watermarked teacher to an initially unwatermarked student of the same architecture. The student retains most of the teacher's detection power at comparable perplexity (Appendix F.7). For multi-key deployment, we examine whether text generated with one key is falsely attributed to another by testing independently sampled mismatched keys. The observed cross-key detection rates remain close to the nominal false-positive level, supporting key-specific attribution in the evaluated setting (Appendix F.2). Finally, Appendix E presents comprehensive ablation studies on how the regularization term, its coefficient, and the quality anchor affect the quality–detectability trade-off, examines sensitivity to GRPO group size and sampling temperature, and evaluates the effects of removing the watermark reward or replacing it with the KGW z-score.

5 Conclusion

We introduced MARKTUNE, a practical and theoretically grounded on-policy fine-tuning framework for watermarking open-weight language models. By optimizing a dual objective that combines a watermark-signal reward derived from the GAUSSMARK test statistic with a quality regularizer that anchors the trained model to a high-quality reference distribution, instantiated either as the cross-entropy on human text or as the KL divergence to a stronger oracle model from the same family, MARKTUNE enables watermark-aware yet quality-preserving exploration in the over-parameterized representation space. Extensive empirical evaluations across models and datasets show that MARKTUNE consistently improves the quality-detectability trade-off of GAUSSMARK, reaching performance close to inference-time watermarking while preserving downstream task performance. Our approach is also very robust to semantic-preserving text-editing and open-weight attacks and generalizes well to unseen data. These results suggest that MARKTUNE offers a flexible and effective framework for embedding reliable, high-quality watermarks into open-weight language models.

Acknowledgements

Z.S.W. was supported in part by an NSF CAREER Award and an NSF Award #2232693.

References

[1] Scott Aaronson and Hendrik Kirchner. Watermarking of large language models. https://www.scottaaronson.com/talks/watermark.ppt, 2022.

[2] Anirudh Ajith, Sameer Singh, and Danish Pruthi. Downstream trade-offs of a family of text watermarks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14039–14053, 2024.

[3] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

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

[5] Tu Bui, Shruti Agarwal, and John Collomosse. TrustMark: Robust watermarking and watermark removal for arbitrary resolution images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18629–18639, 2025.

[6] Clément L Canonne. A short note on an inequality between KL and TV. *arXiv preprint arXiv:2202.07198*, 2022.

[7] Guangyu Chen, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei. WavMark: Watermarking for audio generation. *arXiv preprint arXiv:2308.12770*, 2023.

[8] Huanran Chen, Yinpeng Dong, Zeming Wei, Yao Huang, Yichi Zhang, Hang Su, and Jun Zhu. Unveiling the basin-like loss landscape in large language models. *arXiv preprint arXiv:2505.17646*, 2025.

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

[10] 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.

[11] Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. RingID: Rethinking Tree-Ring watermarking for enhanced multi-key identification. In *European Conference on Computer Vision*, pages 338–354, 2024.

[12] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[13] 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.

[14] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long Form Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, 2019.

[15] Liancheng Fang, Aiwei Liu, Henry Peng Zou, Yankai Chen, Hengrui Zhang, Zhongfen Deng, and Philip Yu. MUSE: Model-agnostic tabular watermarking via multi-sample selection. In *International Conference on Learning Representations*, pages 16541–16574, 2026.

[16] Hanqi Feng, Peng Qiu, Meng-Chun Zhang, You Fan, Yiran Tao, and Barnabas Poczos. Soft metropolis-hastings correction for generative model sampling. *bioRxiv*, 2025.

[17] Hanqi Feng, Peng Qiu, Mengchun Zhang, Yiran Tao, You Fan, Jingtao Xu, and Barnabas Poczos. Learning from b cell evolution: Adaptive multi-expert diffusion for antibody design via online optimization. *arXiv preprint arXiv:2508.02834*, 2025.

[18] Weitao Feng, Wenbo Zhou, Jiyan He, Jie Zhang, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu. AquaLoRA: Toward white-box protection for customized Stable Diffusion models via watermark LoRA. In *International Conference on Machine Learning*, pages 13423–13444, 2024.

[19] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The Stable Signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023.

[20] Pierre Fernandez, Hady Elsahar, I. Zeki Yalniz, and Alexandre Mourachko. Video Seal: Open and efficient video watermarking. *arXiv preprint arXiv:2412.09492*, 2024.

[21] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

[22] Dylan J Foster, Zakaria Mhammedi, and Dhruv Rohatgi. Is a good foundation necessary for efficient reinforcement learning? the computational role of the base model in exploration. *arXiv preprint arXiv:2503.07453*, 2025.

[23] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. *Zenodo*, 2021.

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

[25] Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. OpenWebText Corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

[26] 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.

[27] 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.

[28] 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.

[29] Hengzhi He, Peiyu Yu, Junpeng Ren, Ying Nian Wu, and Guang Cheng. Watermarking generative tabular data. *arXiv preprint arXiv:2405.14018v1*, 2024.

[30] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[31] Runyi Hu, Jie Zhang, Yiming Li, Jiwei Li, Qing Guo, Han Qiu, and Tianwei Zhang. VideoShield: Regulating diffusion-based video generation models via watermarking. In *International Conference on Learning Representations*, pages 52395–52425, 2025.

[32] Xuming Hu, Hanqian Li, Jungang Li, Yu Huang, Shuliang Liu, Qi Zheng, Junhao Chen, and Aiwei Liu. VideoMark: A distortion-free robust watermarking framework for video diffusion models. *arXiv preprint arXiv:2504.16359*, 2025.

[33] 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.

[34] Audrey Huang, Adam Block, Dylan J Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. In *The Thirteenth International Conference on Learning Representations*, 2025.

[35] 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.

[36] Jiahao Huo et al. PMark: Towards robust and distortion-free semantic-level watermarking with channel constraints. In *International Conference on Learning Representations*, 2026.

[37] 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.

[38] 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.

[39] Zhengyuan Jiang, Moyang Guo, Yuepeng Hu, Jinyuan Jia, and Neil Zhenqiang Gong. Certifiably robust image watermark. In *European Conference on Computer Vision*, pages 427–443, 2024.

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

[41] 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.

[42] 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.

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

[44] Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust Distortion-free Watermarks for Language Models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

[45] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

[46] 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.

[47] Chang Liu, Jie Zhang, Tianwei Zhang, Xi Yang, Weiming Zhang, and Nenghai Yu. Detecting voice cloning attacks via timbre watermarking. In *Network and Distributed System Security Symposium*, 2024.

[48] Xiyang Luo, Yinxiao Li, Huiwen Chang, Ce Liu, Peyman Milanfar, and Feng Yang. DVMark: A deep multiscale network for video watermarking. *IEEE Transactions on Image Processing*, 2023.

[49] 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.

[50] George A Miller. Wordnet: a lexical database for English. *Communications of the ACM*, 38 (11):39–41, 1995.

[51] Yurii Nesterov. *Lectures on Convex Optimization*. Springer, 2 edition, 2018.

[52] 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.

[53] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in neural information processing systems*, 2022.

[54] Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. MarkLLM: An open-source toolkit for LLM watermarking. *arXiv preprint arXiv:2405.10051*, 2024.

[55] Qi Pang, Shengyuan Hu, Wenting Zheng, and Virginia Smith. No free lunch in llm watermarking: Trade-offs in watermarking design choices. In *Advances in Neural Information Processing Systems*, volume 37, pages 138756–138788, 2024.

[56] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*, 2021.

[57] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in neural information processing systems*, 2023.

[58] 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.

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

[60] Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre Défossez, Teddy Furon, and Tuan Tran. Proactive detection of voice cloning with localized watermarking. In *International Conference on Machine Learning*, pages 43180–43196, 2024.

[61] Robin San Roman, Pierre Fernandez, Antoine Deleforge, Yossi Adi, and Romain Serizel. Latent watermarking of audio generative models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025.

[62] Tom Sander, Pierre Fernandez, Alain Oliviero Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. In *International Conference on Learning Representations*, pages 79569–79599, 2025.

[63] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[64] 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.

[65] Mayank Kumar Singh, Naoya Takahashi, Weihsiang Liao, and Yuki Mitsufuji. SilentCipher: Deep audio watermarking. In *Interspeech 2024*, pages 2235–2239, 2024.

[66] Chris Stokel-Walker. AI bot ChatGPT writes smart essays-should professors worry? *Nature News*, 2022.

[67] 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.

[68] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[69] 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.

[70] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-Rings watermarks: Invisible fingerprints for diffusion images. In *Advances in Neural Information Processing Systems*, volume 36, pages 58047–58063, 2023.

[71] Yihan Wu, Zhengmian Hu, Hongyang Zhang, and Heng Huang. A resilient and accessible distribution-preserving watermark for large language models. *arXiv preprint arXiv:2310.07710*, 2023.

[72] 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.

[73] Yangxinyu Xie, Xiang Li, Tanwi Mallick, Weijie Su, and Ruixun Zhang. Debiasing watermarks for large language models via maximal coupling. *Journal of the American Statistical Association*, 120(551):1424–1436, 2025.

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

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

[76] 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.

[77] Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian Shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12162–12171, 2024.

[78] 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.

[79] Yulin Zhang, Jiangqun Ni, Wenkang Su, and Xin Liao. A novel deep video watermarking framework with enhanced robustness to H.264/AVC compression. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8095–8104, 2023.

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

[81] Xuandong Zhao, Lei Li, and Yu-Xiang Wang. Permute-and-Flip: An optimally stable and watermarkable decoder for LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025.

[82] Yizhou Zhao, Xiang Li, Peter XK Song, Qi Long, and Weijie Su. Robust spectral watermark for synthetic tabular data. *Statistical Learning and Data Science*, 1:100001, 2026. doi: 10.1016/j.slads.2026.100001.

[83] Yihao Zheng, Haocheng Xia, Junyuan Pang, Jinfei Liu, Kui Ren, Lingyang Chu, Yang Cao, and Li Xiong. TabularMark: Watermarking tabular datasets for machine learning. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 3570–3584, 2024.

[84] Chaoyi Zhu, Jiayi Tang, Jeroen Galjaard, Pin-Yu Chen, Robert Birke, Cornelis Bos, and Lydia Chen. TabWak: A watermark for tabular diffusion models. In *International Conference on Learning Representations*, pages 42174–42201, 2025.

Contents

1 Introduction **1**

1.1 Related Work 3

2 Preliminaries **3**

3 MARKTUNE: On-Policy Fine-Tuning with Watermark Signal Rewards **4**

4 Experiments **6**

4.1 Experimental Setup 7

4.2 Main Results 7

4.3 Robustness Evaluation 9

5 Conclusion **10**

References **10**

A Practical Implementation of MARKTUNE **18**

B Theorems and Proofs **18**

B.1 Proof of Proposition 3.1 18

B.2 Proof of Proposition 3.3 19

B.3 Detectability Guarantee for the GAUSSMARK test 20

B.4 Power of a reverse-KL population policy 22

C A Stylized Linear-Softmax Analysis of MARKTUNE **23**

D Implementation Details **27**

D.1 Evaluation metrics 27

D.2 Watermarking methods 28

D.3 Fine-tuning attack 28

D.4 Computation time benchmark 28

D.5 Detection access and memory 28

D.6 Computational Cost 29

E Ablation Study **29**

E.1 Gaussian initialization 29

E.2 Regularization term and coefficient 30

E.3 Choice of quality anchor 30

E.4 Removing the watermark reward 31

E.5 GRPO sampling hyperparameters 32

E.6 Alternative watermark rewards 32

E.7 Choice of watermarked parameters 32

F Supplemental Results **33**
F.1 Detection at ultra low false-positive rates . . . 33
F.2 Multi-key deployment . . . 34
F.3 Additional model-embedded baseline . . . 34
F.4 Decoding strategy and output length . . . 34
F.5 Robustness to text editing . . . 35
F.6 Quantization, pruning, and model merging . . . 36
F.7 Robustness to logit-based distillation . . . 36
F.8 Watermark reward during training . . . 37

G Limitations and Future Directions **37**

H Examples of Generated Texts **38**
H.1 Example 1 . . . 38
H.2 Example 2 . . . 39
H.3 Example 3 . . . 41

I AI Use Statement **44**

J Ethics Statement **44**

A Practical Implementation of MARKTUNE

Algorithm 2 GRPO for MARKTUNE

Input: initial model $p_{\theta(\xi_\sigma)}$; watermark signal reward $\mathcal{R}_{\text{wm}}(\cdot;\xi_\sigma)$; text dataset $\mathcal{D}$; hyperparameters $\varepsilon, \lambda, T_1, T_2$.
Initialize: Set $p_\theta \leftarrow p_{\theta(\xi_\sigma)}$.
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 objective Eq.(5).
  end for
end for
Output: p_θ.

To optimize the objective in Eq.(3), we use the GRPO [64] algorithm. This choice is motivated by the fact that Eq.(3) cannot be directly optimized due to its dependence on sampled outputs $y \sim p_{\theta(\xi_\sigma)}$. 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 [17].

For each prompt from a batch $x \sim \mathcal{D}_b$, we sample a group of responses $\{y_j\}_{j=1}^G$ from the current policy 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 $\rho_j = \frac{p_\theta(y_j|x)}{p_{\theta_{\text{old}}}(y_j|x)}$, maximizing $\mathbb{E}\Big[\min\big(\rho_j \widehat{A}_j,\ \text{clip}(\rho_j, 1-\varepsilon, 1+\varepsilon)\widehat{A}_j\big)\Big]$ together with a specified language modeling regularization term. The overall GRPO objective is given by

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

See Algorithm 2 for the detailed implementation of GRPO.

B Theorems and Proofs

B.1 Proof of Proposition 3.1

Proof. Fix a prompt x and abbreviate $p_\theta := p_\theta(\cdot \mid x)$. Define

$$F : \mathbb{R}^{d_r} \to \mathbb{R}, \qquad F(\Delta) \ := \ \text{D}_{\text{KL}}\big(p_{(\theta_{\text{wm}}+\Delta,\,\theta_0)}(\cdot \mid x)\,\big\|\,p_\theta\big)\,.$$

The proof proceeds in three steps: (i) verify $F(0) = 0$ and $\nabla F(0) = 0$; (ii) use a uniform bound on the Hessian of F to derive a pointwise quadratic upper bound on F; (iii) average over $\xi_\sigma \sim \mathcal{N}(0, \sigma^2\mathbb{I}_{d_r})$. A short corollary translates the resulting KL bound into an expected-TV bound.

Step 1: $F(0) = 0$ and $\nabla F(0) = 0$. At $\Delta = 0$, the perturbed parameters coincide with the base parameters, so $p_{(\theta_{\mathrm{wm}}+\Delta,\theta_0)} = p_\theta$ and $F(0) = \mathrm{D}_{\mathrm{KL}}(p_\theta \,\|\, p_\theta) = 0$. The KL divergence is nonnegative, so $\Delta = 0$ is a global minimizer of F. The differentiability assumption together with the first-order optimality condition then gives $\nabla F(0) = 0$.

Step 2: pointwise quadratic upper bound. By assumption, F is twice continuously differentiable with $\sup_{\Delta \in \mathbb{R}^{d_r}} \|\nabla^2 F(\Delta)\|_{\mathrm{op}} \le M$ for some constant $M > 0$. Taylor's theorem with integral remainder, combined with $F(0) = 0$ and $\nabla F(0) = 0$, yields

$$F(\Delta) \;=\; \int_0^1 (1-t)\, \Delta^\top \nabla^2 F(t\Delta)\, \Delta \, dt. \tag{6}$$

For every $t \in [0,1]$, the operator-norm bound on the Hessian gives $\Delta^\top \nabla^2 F(t\Delta)\, \Delta \le \|\nabla^2 F(t\Delta)\|_{\mathrm{op}} \|\Delta\|^2 \le M \|\Delta\|^2$. Substituting this estimate into (6) and using $\int_0^1 (1-t)\, dt = 1/2$,

$$F(\Delta) \;\le\; \frac{M}{2} \|\Delta\|^2, \qquad \forall\, \Delta \in \mathbb{R}^{d_r}. \tag{7}$$

Step 3: averaging over ξ_σ. Specializing $\Delta = \xi_\sigma$ in (7) and taking expectation over $\xi_\sigma \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_r})$, we use $\mathbb{E}\,\|\xi_\sigma\|^2 = \sigma^2 d_r$ to obtain

$$\mathbb{E}_{\xi_\sigma}\big[\mathrm{D}_{\mathrm{KL}}\big(p_{\theta(\xi_\sigma)}(\cdot \mid x) \,\big\|\, p_\theta\big)\big] \;=\; \mathbb{E}_{\xi_\sigma}[F(\xi_\sigma)] \;\le\; \frac{M}{2}\, \mathbb{E}\,\|\xi_\sigma\|^2 \;=\; \frac{M}{2}\, \sigma^2 d_r,$$

which establishes the bound with $C = M/2$.

Remark on the Hessian bound. The constant M admits a local-curvature interpretation. Differentiating F twice and evaluating at $\Delta = 0$ recovers the Fisher information matrix at θ_{wm}, $\nabla^2 F(0) = \mathcal{I}(\theta_{\mathrm{wm}})$, with entries

$$\mathcal{I}(\theta_{\mathrm{wm}})_{jk} \;=\; \mathbb{E}_{y \sim p_\theta}\left[\frac{\partial \log p_\theta(y)}{\partial \theta^{(j)}_{\mathrm{wm}}}\, \frac{\partial \log p_\theta(y)}{\partial \theta^{(k)}_{\mathrm{wm}}}\right].$$

Hence M captures the local Lipschitz sensitivity of the map $\theta_{\mathrm{wm}} \mapsto p_{(\theta_{\mathrm{wm}},\theta_0)}(\cdot \mid x)$. In the worst case M scales with a network Lipschitz factor (e.g. products of layer operator norms), which can in principle grow exponentially in depth; empirically [4] the effective scaling is much milder, especially when the watermarked subspace θ_{wm} is restricted to later layers.

Corollary: bound in total variation. Pinsker's inequality [6] gives, for every realization of ξ_σ, $\mathrm{TV}\big(p_{\theta(\xi_\sigma)},\, p_\theta\big)^2 \le \frac{1}{2}\, \mathrm{D}_{\mathrm{KL}}\big(p_{\theta(\xi_\sigma)} \,\|\, p_\theta\big)$. Combining with Jensen's inequality applied to the concave function $\sqrt{\cdot}$,

$$\mathbb{E}_{\xi_\sigma}\big[\mathrm{TV}\big(p_{\theta(\xi_\sigma)},\, p_\theta\big)\big] \;\le\; \sqrt{\tfrac{1}{2}\, \mathbb{E}_{\xi_\sigma}\big[\mathrm{D}_{\mathrm{KL}}\big(p_{\theta(\xi_\sigma)} \,\|\, p_\theta\big)\big]} \;\le\; \sigma \sqrt{\tfrac{C\, d_r}{2}}\,,$$

so the expected total variation distortion grows only as $O(\sigma\sqrt{d_r})$.

□

B.2 Proof of Proposition 3.3

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

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

where the last line is based on

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

for any finite, nonzero vector $\nabla_{\theta_{\mathrm{wm}}} \log p_{\theta_{\mathrm{ref}}}$. The last equality is derived by plugging in $\tau_\alpha = \Phi^{-1}(1-\alpha)$. □

One-sided p-values. For a finite, nonzero reference gradient, the same argument gives

$$p(y,\xi_\sigma) = 1 - \Phi(\psi(y,\xi_\sigma)). \tag{8}$$

For every fixed text y independent of ξ_σ, the statistic is standard normal over the Gaussian key, so $p(y,\xi_\sigma)$ is uniform on $[0,1]$ under the null. Fine-tuning changes the alternative distribution but leaves this null argument unchanged because the clean reference and detection statistic are fixed. Appendix F.1 reports the corresponding ultra-low-FPR evaluation.

B.3 Detectability Guarantee for the GAUSSMARK test

Proposition 3.1 controls the quality cost of GAUSSMARK but does not, on its own, certify detectability. Because MARKTUNE uses the same detection statistic as GAUSSMARK, it retains the key-randomized null law; detection power also depends on the generating distribution and does not transfer automatically to a fine-tuned model. The detection-power analysis of Block et al. [4] already supplies such a guarantee against the full null class for the directly perturbed model. For completeness and to make the dependence on σ and d_r explicit, we record below a stylized Gaussian illustration and a formal exponential-concentration theorem. Appendix B.4 treats the optimized reverse-KL population policy separately.

Proposition B.1 (Stylized Gaussian illustration)**.** *Fix a prompt x and consider a Gaussian model with $p_\theta(\cdot \mid x) = \mathcal{N}(\theta_{\mathrm{wm}}, \mathbb{I}_{d_r})$, so that $\nabla_{\theta_{\mathrm{wm}}} \log p_\theta(y \mid x) = y - \theta_{\mathrm{wm}}$. Let $\xi_\sigma \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_r})$ and let $\psi(y, \xi_\sigma \mid x)$ denote the test statistic of (1). Under $\mathbf{H_A}$ where $y \sim \mathcal{N}(\theta_{\mathrm{wm}} + \xi_\sigma, \mathbb{I}_{d_r})$, there exist absolute constants $c_0, c_1 > 0$ such that, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\psi(y, \xi_\sigma \mid x) \;\geq\; c_0 \frac{\sigma\sqrt{d_r}}{\sqrt{1+\sigma^2}} \qquad \textit{whenever} \qquad \frac{\sigma^2 d_r}{1+\sigma^2} \;\geq\; c_1 \log(6/\delta).$$

Proof. Write $y = \theta_{\mathrm{wm}} + \xi_\sigma + z$ with $z \sim \mathcal{N}(0, \mathbb{I}_{d_r})$ independent of ξ_σ. Since $\nabla_{\theta_{\mathrm{wm}}} \log p_\theta(y \mid x) = y - \theta_{\mathrm{wm}} = \xi_\sigma + z$, the test statistic of (1) expands as

$$\psi(y, \xi_\sigma \mid x) \;=\; \frac{\langle \xi_\sigma,\, \xi_\sigma + z \rangle}{\sigma\, \|\xi_\sigma + z\|_2} \;=\; \frac{\|\xi_\sigma\|_2^2 + \langle \xi_\sigma,\, z \rangle}{\sigma\, \|\xi_\sigma + z\|_2}. \tag{9}$$

Set $t := \log(6/\delta)$. We control the numerator from below and the denominator from above on three high-probability events whose intersection holds with probability at least $1-\delta$.

Step 1 (lower bound on $\|\xi_\sigma\|_2^2$). Since $\|\xi_\sigma\|_2^2/\sigma^2 \sim \chi^2_{d_r}$, the chi-square deviation inequality [45, Lemma 1] gives, with probability at least $1 - e^{-t} = 1 - \delta/6$,

$$\|\xi_\sigma\|_2^2 \;\geq\; \sigma^2 \left(d_r - 2\sqrt{d_r\, t}\right).$$

Step 2 (upper bound on $|\langle \xi_\sigma, z \rangle|$). Conditional on ξ_σ, $\langle \xi_\sigma, z \rangle \sim \mathcal{N}(0,\, \|\xi_\sigma\|_2^2)$ since $z \sim \mathcal{N}(0, \mathbb{I}_{d_r})$ is independent of ξ_σ. The Gaussian tail bound gives $\Pr(|\langle \xi_\sigma, z \rangle| > \|\xi_\sigma\|_2 \sqrt{2t} \mid \xi_\sigma) \leq 2e^{-t} = \delta/3$. Combined with the chi-square upper deviation $\|\xi_\sigma\|_2^2 \leq \sigma^2(d_r + 2\sqrt{d_r t} + 2t)$ [45, Lemma 1] (which in turn holds with probability at least $1 - e^{-t}$), an absorbed-constant calculation gives, with probability at least $1 - \delta/2$,

$$|\langle \xi_\sigma,\, z \rangle| \;\leq\; \|\xi_\sigma\|_2 \sqrt{2t} \;\leq\; C_1\, \sigma \left(\sqrt{d_r\, t} + t\right),$$

where $C_1 > 0$ is an absolute constant.

Step 3 (upper bound on $\|\xi_\sigma + z\|_2$). Since $\xi_\sigma + z \sim \mathcal{N}(0,\, (1+\sigma^2)\, \mathbb{I}_{d_r})$, $\|\xi_\sigma + z\|_2^2/(1+\sigma^2) \sim \chi^2_{d_r}$. Chi-square upper concentration [45, Lemma 1] gives, with probability at least $1 - \delta/6$,

$$\|\xi_\sigma + z\|_2 \;\leq\; \sqrt{(1+\sigma^2)\,(d_r + 2\sqrt{d_r\, t} + 2t)} \;\leq\; C_2 \sqrt{1+\sigma^2}\, \left(\sqrt{d_r} + \sqrt{t}\right),$$

for an absolute constant $C_2 > 0$.

Combining. On the intersection of these events (whose failure probability is at most $\delta/6+\delta/2+\delta/6 < \delta$), substituting the three bounds into (9),

$$\psi(y, \xi_\sigma \mid x) \;\geq\; \frac{\sigma^2\,(d_r - 2\sqrt{d_r\,t}) - C_1\,\sigma\,(\sqrt{d_r\,t} + t)}{\sigma\,C_2\,\sqrt{1+\sigma^2}\,(\sqrt{d_r} + \sqrt{t})} \;=\; \frac{\sigma\,\sqrt{d_r}}{\sqrt{1+\sigma^2}} \cdot \mathcal{R}(\sigma, d_r, t),$$

where $\mathcal{R}(\sigma, d_r, t)$ collects the residual ratio. Under the assumption $\sigma^2 d_r/(1+\sigma^2) \geq c_1\,\log(6/\delta)$ for $c_1 > 0$ sufficiently large, elementary algebra shows that $\mathcal{R}(\sigma, d_r, t) \geq c_0$ for some absolute constant $c_0 > 0$. This yields

$$\psi(y, \xi_\sigma \mid x) \;\geq\; c_0\,\frac{\sigma\,\sqrt{d_r}}{\sqrt{1+\sigma^2}},$$

as claimed. The constants c_0, c_1, C_1, C_2 depend only on the chi-square and Gaussian tail constants of Laurent and Massart [45]. □

The proposition shows that, even in this idealized regime, the watermark statistic grows like $\Omega(\sigma\sqrt{d_r}/\sqrt{1+\sigma^2})$ under the alternative, while remaining $\mathcal{N}(0,1)$ under the null by Proposition 3.3. The next theorem promotes this intuition to an explicit power guarantee against the testing alternative, holding in general (not necessarily Gaussian) language models under a strong-concavity assumption on the log-likelihood in θ_{wm}.

Theorem B.2 (Power of the level-α GAUSSMARK test)**.** *Let $\alpha \in (0,1)$ and $\tau_\alpha = \Phi^{-1}(1-\alpha)$ be as in Proposition 3.3, and let ϕ_α denote the test that rejects $\mathbf{H_0}$ when $\psi(y, \xi_\sigma \mid x) \geq \tau_\alpha$. Assume that, for every response y and every prompt x, the map $\theta_{\mathrm{wm}} \mapsto \log p_{(\theta_{\mathrm{wm}}, \theta_0)}(y \mid x)$ is η-strongly concave on $\mathbb{R}^{d_r}$. Then ϕ_α has power at least $1-\beta$ under $\mathbf{H_A}$ whenever*

$$d_r \;\geq\; \frac{2\log\,\mathbb{E}_{y\sim p_\theta(\cdot|x)}\Big[\exp\big(\tau_\alpha\,\sigma\,\|\nabla_{\theta_{\mathrm{wm}}}\log p_\theta(y \mid x)\|_2\big)\Big] + 2\log(1/\beta)}{\log(1+\eta\sigma^2)}.$$

Proof. Throughout the proof we condition on the prompt x and abbreviate $p_\theta := p_\theta(\cdot \mid x)$, $g_\theta(y) := \nabla_{\theta_{\mathrm{wm}}} \log p_\theta(y)$, and $\theta(\xi_\sigma) := (\theta_{\mathrm{wm}} + \xi_\sigma, \theta_0)$. The test ϕ_α rejects $\mathbf{H_0}$ when $\psi(y, \xi_\sigma \mid x) \geq \tau_\alpha$, so its type-II error under $\mathbf{H_A}$ (where $y \sim p_{\theta(\xi_\sigma)}$ and $\xi_\sigma \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_r})$) is

$$\beta(\phi_\alpha) \;:=\; \Pr_{\mathbf{H_A}}\big(\psi(y, \xi_\sigma \mid x) < \tau_\alpha\big).$$

Step 1: change of measure to the base model. Wherever $p_\theta(y) > 0$, the likelihood ratio $\Lambda(y, \xi_\sigma) := p_{\theta(\xi_\sigma)}(y)/p_\theta(y)$ is well-defined. The standard change-of-measure identity gives

$$\beta(\phi_\alpha) \;=\; \mathbb{E}_{\xi_\sigma}\big[\,\mathbb{E}_{y\sim p_\theta}\big[\Lambda(y, \xi_\sigma)\,\mathbf{1}\{\psi(y, \xi_\sigma \mid x) < \tau_\alpha\}\big]\big]\,, \tag{10}$$

where in the new measure y and ξ_σ are independent (the dominating measure is $p_\theta \otimes \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_r})$).

Step 2: bound Λ on the event $\{\psi < \tau_\alpha\}$. By assumption, the map $\theta'_{\mathrm{wm}} \mapsto \log p_{(\theta'_{\mathrm{wm}}, \theta_0)}(y)$ is η-strongly concave on $\mathbb{R}^{d_r}$ for every fixed y. The standard quadratic upper bound for η-strongly concave functions [51, Theorem 2.1.10] gives, for any $\xi \in \mathbb{R}^{d_r}$,

$$\log p_{\theta(\xi)}(y) - \log p_\theta(y) \;\leq\; \langle \xi,\, g_\theta(y)\rangle \;-\; \tfrac{\eta}{2}\,\|\xi\|_2^2. \tag{11}$$

On the event $\{\psi(y, \xi_\sigma) < \tau_\alpha\}$, the definition of ψ in (1) implies $\langle \xi_\sigma, g_\theta(y)\rangle < \tau_\alpha\,\sigma\,\|g_\theta(y)\|_2$. Combining this with (11) at $\xi = \xi_\sigma$ and exponentiating,

$$\Lambda(y, \xi_\sigma) \;\leq\; \exp\Big(\tau_\alpha\,\sigma\,\|g_\theta(y)\|_2 \;-\; \tfrac{\eta}{2}\,\|\xi_\sigma\|_2^2\Big) \quad \text{on } \{\psi(y, \xi_\sigma) < \tau_\alpha\}. \tag{12}$$

Step 3: integrate out ξ_σ. Substituting (12) into (10) and dropping the indicator (which is at most one),

$$\beta(\phi_\alpha) \;\leq\; \mathbb{E}_{\xi_\sigma}\big[\mathbb{E}_{y\sim p_\theta}\big[\exp\big(\tau_\alpha\,\sigma\,\|g_\theta(y)\|_2 - \tfrac{\eta}{2}\,\|\xi_\sigma\|_2^2\big)\big]\big]\,.$$

Under the dominating product measure $p_\theta \otimes \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_r})$, the random variables y and ξ_σ are independent, so Fubini's theorem factorizes the expectation:

$$\beta(\phi_\alpha) \;\le\; \mathbb{E}_{\xi_\sigma}\left[e^{-\eta\,\|\xi_\sigma\|_2^2/2}\right] \cdot \mathbb{E}_{y\sim p_\theta}\left[e^{\tau_\alpha\,\sigma\,\|g_\theta(y)\|_2}\right]. \tag{13}$$

For the Gaussian moment, since the components of ξ_σ are independent $\mathcal{N}(0, \sigma^2)$ random variables and $\eta\,\sigma^2 > -1$,

$$\mathbb{E}_{\xi_\sigma}\left[e^{-\eta\,\|\xi_\sigma\|_2^2/2}\right] \;=\; \prod_{j=1}^{d_r} \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[e^{-\eta\,\sigma^2\,Z^2/2}\right] \;=\; (1+\eta\,\sigma^2)^{-d_r/2},$$

where the last identity uses the standard one-dimensional moment generating function of a chi-square: $\mathbb{E}[e^{-\lambda Z^2/2}] = (1+\lambda)^{-1/2}$ for $\lambda > -1$. Substituting back into (13),

$$\beta(\phi_\alpha) \;\le\; (1+\eta\,\sigma^2)^{-d_r/2} \cdot \mathbb{E}_{y\sim p_\theta}\left[\exp\big(\tau_\alpha\,\sigma\,\|g_\theta(y)\|_2\big)\right]. \tag{14}$$

Step 4: solve for d_r. Imposing $\beta(\phi_\alpha) \le \beta$ in (14), taking logarithms, and rearranging gives

$$\frac{d_r}{2}\,\log(1+\eta\,\sigma^2) \;\ge\; \log \mathbb{E}_{y\sim p_\theta}\left[\exp\big(\tau_\alpha\,\sigma\,\|g_\theta(y)\|_2\big)\right] \;+\; \log(1/\beta),$$

which is equivalent to the lower bound on d_r stated in the theorem. □

Theorem B.2 makes explicit the role of σ and d_r on the detectability side: the type-II error decays as $(1+\eta\sigma^2)^{-d_r/2}$, while Proposition 3.1 shows the quality cost grows only as $\sigma^2 d_r$.

B.4 Power of a reverse-KL population policy

The GAUSSMARK power guarantee does not automatically transfer to a model after GRPO fine-tuning. We instead analyze the converged population optimizer of the reverse-KL objective in the linear-softmax setting. This gives a detection-power guarantee for the optimized policy under the assumptions below.

Theorem B.3 (Detection power of a reverse-KL MARKTUNE population policy)**.** *Fix a prompt x and suppress it from the notation. Consider the linear-softmax model*

$$p_\theta(y) \propto \exp\{\langle \theta_{\mathrm{wm}}, \Phi_{\mathrm{wm}}(y)\rangle + \langle \theta_0, \Phi_0(y)\rangle\},$$

with a fixed feature map Φ. Let $p_{\theta_{\mathrm{ref}}}$ be the clean reference used for detection, $q = p_{\theta^\star}$ the regularization distribution, and

$$g_{\mathrm{ref}}(y) = \nabla_{\theta_{\mathrm{wm}}} \log p_{\theta_{\mathrm{ref}}}(y), \qquad b = \mathbb{E}_{p_{\theta_{\mathrm{ref}}}}[\Phi_{\mathrm{wm}}(y)] - \mathbb{E}_q[\Phi_{\mathrm{wm}}(y)].$$

Draw $\xi_\sigma \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d_r})$, and let $\pi^\star_{\xi_\sigma}$ maximize the population objective

$$J_{\xi_\sigma}(\pi) = \mathbb{E}_{y\sim\pi}[\langle \xi_\sigma, g_{\mathrm{ref}}(y)\rangle] - \lambda D_{\mathrm{KL}}(\pi\|q), \qquad \lambda > 0. \tag{15}$$

Assume finite, nonzero reference gradients and that $\theta_{\mathrm{wm}} \mapsto \log p_{(\theta_{\mathrm{wm}}, \theta_0^\star)}(y)$ is η-strongly concave for every y. Assume the normalizing constants and displayed moments are finite. Under the alternative $y \mid \xi_\sigma \sim \pi^\star_{\xi_\sigma}$, the level-$\alpha$ test has power at least $1-\beta$ whenever

$$d_r \ge \frac{2\log \mathbb{E}_{y\sim q}\left[\exp\left\{\frac{\tau_\alpha \sigma}{\lambda}\|g_{\mathrm{ref}}(y)\|_2\right\}\right] + \dfrac{\sigma^2\|b\|_2^2}{\lambda^2+\eta\sigma^2} + 2\log(1/\beta)}{\log\left(1+\dfrac{\eta\sigma^2}{\lambda^2}\right)}. \tag{16}$$

Proof. Write $r_{\xi_\sigma}(y) = \langle \xi_\sigma, g_{\mathrm{ref}}(y)\rangle$ and

$$Z_{\xi_\sigma} = \mathbb{E}_{y\sim q} \exp\left\{\lambda^{-1} r_{\xi_\sigma}(y)\right\}.$$

For any $\pi \ll q$ with finite reward expectation and KL divergence, the exponentially tilted distribution satisfies

$$D_{\mathrm{KL}}\big(\pi\|\pi^\star_{\xi_\sigma}\big) = D_{\mathrm{KL}}(\pi\|q) - \lambda^{-1}\mathbb{E}_\pi[r_{\xi_\sigma}(y)] + \log Z_{\xi_\sigma}.$$

Thus $J_{\xi_\sigma}(\pi) = \lambda \log Z_{\xi_\sigma} - \lambda D_{\mathrm{KL}}(\pi \| \pi^\star_{\xi_\sigma})$, and the maximizing distribution is

$$\pi^\star_{\xi_\sigma}(y) = \frac{q(y)\exp\{\lambda^{-1} r_{\xi_\sigma}(y)\}}{Z_{\xi_\sigma}} = p_{(\theta^\star_{\mathrm{wm}} + \lambda^{-1}\xi_\sigma, \theta^\star_0)}(y). \tag{17}$$

The second equality follows from the linear-softmax score identity $g_{\mathrm{ref}}(y) = \Phi_{\mathrm{wm}}(y) - \mathbb{E}_{p_{\theta_{\mathrm{ref}}}}[\Phi_{\mathrm{wm}}(y)]$: the reference expectation is constant in y and is absorbed into the normalizer.

Let $\Lambda(y, \xi_\sigma) = \pi^\star_{\xi_\sigma}(y)/q(y)$. Changing measure to q expresses the Type-II error as

$$\beta(\phi_\alpha) = \mathbb{E}_{\xi_\sigma}\mathbb{E}_{y\sim q}[\Lambda(y,\xi_\sigma)\mathbf{1}\{\psi(y,\xi_\sigma) < \tau_\alpha\}]\,.$$

By strong concavity and the identity $\nabla_{\theta_{\mathrm{wm}}} \log q(y) = g_{\mathrm{ref}}(y) + b$,

$$\log \Lambda(y,\xi_\sigma) \le \lambda^{-1}\langle \xi_\sigma, g_{\mathrm{ref}}(y) + b\rangle - \frac{\eta}{2\lambda^2}\|\xi_\sigma\|_2^2.$$

On the failure event $\{\psi < \tau_\alpha\}$, this gives

$$\log \Lambda(y,\xi_\sigma) \le \frac{\tau_\alpha \sigma}{\lambda}\|g_{\mathrm{ref}}(y)\|_2 + \frac{\langle \xi_\sigma, b\rangle}{\lambda} - \frac{\eta}{2\lambda^2}\|\xi_\sigma\|_2^2.$$

Dropping the indicator and using independence under the dominating product measure,

$$\begin{aligned}
\beta(\phi_\alpha) &\le \mathbb{E}_q \exp\Big\{\frac{\tau_\alpha\sigma}{\lambda}\|g_{\mathrm{ref}}(y)\|_2\Big\}\,\mathbb{E}_{\xi_\sigma}\exp\Big\{\frac{\langle \xi_\sigma, b\rangle}{\lambda} - \frac{\eta\|\xi_\sigma\|_2^2}{2\lambda^2}\Big\} \\
&= \mathbb{E}_q \exp\Big\{\frac{\tau_\alpha\sigma}{\lambda}\|g_{\mathrm{ref}}(y)\|_2\Big\}\left(1 + \frac{\eta\sigma^2}{\lambda^2}\right)^{-d_r/2}\exp\Big\{\frac{\sigma^2\|b\|_2^2}{2(\lambda^2 + \eta\sigma^2)}\Big\}\,.
\end{aligned}$$

The last equality follows by completing the square in the Gaussian integral. Requiring this bound to be at most β, taking logarithms, and rearranging yields Eq. (16). □

Scope of the result. The theorem concerns the converged reverse-KL population policy with an unnormalized alignment reward. Its strong-concavity condition is an explicit idealization, not a property established for the Transformer models used in our experiments or a consequence of the linear-softmax parameterization alone. The practical objective uses normalized rewards and CE regularization, so the exponential-tilt argument is not a literal characterization of the finite-step GRPO checkpoint. The local CE analysis in Appendix C provides a complementary explanation of the quality–reward trade-off; detection power of the trained LLMs is evaluated empirically.

C A Stylized Linear-Softmax Analysis of MARKTUNE

In this section we provide a stylized analysis of how MARKTUNE can improve the quality-detectability trade-off. The goal is not to reproduce all realistic training dynamics, but rather to study the local geometry of the objective optimized in the main text, within an idealized linear-softmax model.

Model and notation. Let $\theta \in \mathbb{R}^d$ denote the full set of model weights. For a fixed watermark key ξ, a weight-editing scheme such as GAUSSMARK produces an initial watermarked weight $\theta(\xi)$ (here we omit ξ's dependence on σ for simplicity); MARKTUNE then fine-tunes $\theta(\xi)$ by optimizing a reward-regularized objective.

For the analysis, we adopt a standard linear-softmax setting:

$$p_\theta(y \mid x) \;\propto\; \exp\big\{\langle \theta,\, \Phi(x,y)\rangle\big\}, \tag{18}$$

where $\Phi(x,y) \in \mathbb{R}^d$ is a fixed feature map that encodes the prompt-response pair.

Let $p^\star(\cdot \mid x)$ denote a high-quality target conditional distribution (e.g., the population language distribution or an oracle model). We assume that $p^\star$ is realized within the linear-softmax family by a weight $\theta^\star$:

$$p^\star(y \mid x) \;=\; p_{\theta^\star}(y \mid x) \;\propto\; \exp\big\{\langle \theta^\star,\, \Phi(x,y)\rangle\big\}.$$

The population cross entropy risk of a weight θ relative to $p^\star$ is

$$\mathcal{L}(\theta) \;:=\; \mathbb{E}_x\,\mathbb{E}_{y\sim p^\star(\cdot|x)}\big[-\log p_\theta(y \mid x)\big]. \tag{19}$$

As in the main text, the watermark reward is induced by the GAUSSMARK test statistic (for stylized analysis we include the prompt x in the test statistic). Let $p_{\theta_{\mathrm{ref}}}$ be the fixed reference model used for detection, and define

$$h(x, y; \xi) \;:=\; \Big\langle \xi,\; \nabla_{\theta_{\mathrm{wm}}} \log p_{\theta_{\mathrm{ref}}}(y \mid x) \Big\rangle, \tag{20}$$

where θ_{wm} denotes the watermarked block of the reference model parameters. The unnormalized watermark reward of θ is then

$$\mathcal{R}(\theta) \;:=\; \mathbb{E}_x\, \mathbb{E}_{y \sim p_\theta(\cdot|x)} \big[h(x, y; \xi) \big]. \tag{21}$$

The stylized MARKTUNE objective in this setting is

$$J(\theta) \;=\; \mathcal{R}(\theta) \;-\; \lambda\, \mathcal{L}(\theta),$$

where $\lambda > 0$ is a regularization coefficient controlling the trade-off between watermark reward and regularization term. For the sake of analysis, we assume that the GAUSSMARK-soft-activated weight $\theta(\xi)$ lies in a neighborhood of $\theta^\star$, and study the local geometry of J around the high-quality target weight $\theta^\star$.

Goal. We want to show that, in a neighborhood of $\theta^\star$, there exists a perturbation $\delta\theta$ in the full parameter space such that the adjusted weight

$$\theta_{\textsc{MarkTune}} \;:=\; \theta^\star + \delta\theta$$

satisfies, for a small scalar $\kappa = 1/\lambda > 0$:

(i) the cross entropy relative to $p^\star$ changes only at *second order*, $\mathcal{L}(\theta_{\textsc{MarkTune}}) - \mathcal{L}(\theta^\star) = \Theta(\kappa^2)$,

(ii) the watermark reward increases at *first order*, $\mathcal{R}(\theta_{\textsc{MarkTune}}) - \mathcal{R}(\theta^\star) = \Theta(\kappa)$,

When this construction uses the normalized reward with a nonzero gradient and the local location-shift assumption specified below holds, it follows that, under $p_{\theta_{\textsc{MarkTune}}}$ the (normalized) detection statistic exhibits a mean shift of order κ, improving detection power at fixed false positive rate, while the degradation in generation quality is quadratically smaller.

Assumptions. We adopt the following standard local assumptions for linear soft-max models.

(A1) *Smoothness.* The map $\theta \mapsto p_\theta$ is smooth in a neighborhood of $\theta^\star$.

(A2) *Local quadratic expansion of $\mathcal{L}$.* Around $\theta^\star$, the cross-entropy admits the second-order expansion

$$\mathcal{L}(\theta^\star + \delta) \;=\; \mathcal{L}(\theta^\star) \;+\; \frac{1}{2}\, \delta^\top F\, \delta \;+\; o\big(\|\delta\|^2\big),$$

where F is the Fisher information matrix at $\theta^\star$:

$$F \;:=\; \mathbb{E}_x\, \mathrm{Cov}_{y \sim p^\star(\cdot|x)}\big[\Phi(x, y)\big] \;\in\; \mathbb{R}^{d \times d},$$

assumed positive definite on the relevant subspace.

(A3) *Local first-order expansion of $\mathcal{R}$.* For small perturbations δ, the reward admits the first-order expansion

$$\mathcal{R}(\theta^\star + \delta) \;=\; \mathcal{R}(\theta^\star) \;+\; g^\top \delta \;+\; o\big(\|\delta\|\big),$$

for some gradient $g \in \mathbb{R}^d$ characterized in Lemma C.2 below under linear softmax setting.

Assumptions (A2)-(A3) are the usual Fisher-geometry expansions of cross entropy and reward around a population optimum in linear soft-max models.

Local dual optimization problem. Based on the assumptions, near $\theta^\star$, the MARKTUNE objective can be approximated by

$$J(\theta^\star + \delta\theta) \;\approx\; \mathcal{R}(\theta^\star) + g^\top \delta\theta \;-\; \lambda\Big(\mathcal{L}(\theta^\star) + \tfrac{1}{2}\,\delta\theta^\top F\,\delta\theta\Big).$$

The local problem becomes

$$\max_{\delta\theta\in\mathbb{R}^d} \;\; g^\top \delta\theta \;-\; \frac{\lambda}{2}\,\delta\theta^\top F\,\delta\theta := J_{\text{loc}}(\delta\theta). \tag{22}$$

This is a strictly concave quadratic program in parameter space with a unique optimizer.

Lemma C.1 (Closed-form optimizer in parameter space)**.** *The unique maximizer of* (22) *is*

$$\delta\theta^\dagger \;=\; \frac{1}{\lambda}\,F^{-1}\,g,$$

and the optimal objective value equals $\frac{1}{2\lambda}\,g^\top F^{-1} g$.

Proof. The objective is $J_{\text{loc}}(\delta\theta) = g^\top\delta\theta - \frac{\lambda}{2}\delta\theta^\top F\,\delta\theta$. Differentiating and setting the gradient to zero gives the first-order condition

$$g - \lambda F\,\delta\theta = 0 \quad\Rightarrow\quad \delta\theta^\dagger = \lambda^{-1}F^{-1}g.$$

Positive definiteness of F implies strict concavity and uniqueness. Substituting $\delta\theta^\dagger$ back into J_{loc} yields $J_{\text{loc}}(\delta\theta^\dagger) = \frac{1}{2\lambda}g^\top F^{-1}g$. □

Lemma C.2 (Reward gradient in the linear-softmax model)**.** *Let*

$$\mathcal{R}(\theta) = \mathbb{E}_x\,\mathbb{E}_{y\sim p_\theta(\cdot|x)}\big[h(x,y;\xi)\big],$$

with $h(x,y;\xi)$ *as in* (20)*. Then the gradient of* $\mathcal{R}$ *at* $\theta^\star$ *is*

$$g \;:=\; \nabla_\theta\mathcal{R}(\theta)\big|_{\theta=\theta^\star} \;=\; \mathbb{E}_x\,\mathbb{E}_{y\sim p^\star(\cdot|x)}\big[h(x,y;\xi)\,s(x,y)\big],$$

where

$$s(x,y) \;:=\; \nabla_\theta \log p_\theta(y\mid x)\big|_{\theta=\theta^\star} \;=\; \Phi(x,y) - \mu(x), \qquad \mu(x) := \mathbb{E}_{y\sim p^\star(\cdot|x)}[\Phi(x,y)].$$

Equivalently,

$$g \;=\; \mathbb{E}_x\,\mathrm{Cov}_{y\sim p^\star(\cdot|x)}\big[h(x,y;\xi),\;\Phi(x,y)\big],$$

this implies that the reward gradient picks out directions along which the watermark statistic co-varies with the feature representation under the target distribution.

Proof. By definition,

$$\mathcal{R}(\theta) = \mathbb{E}_x\,\mathbb{E}_{y\sim p_\theta(\cdot|x)}[h(x,y;\xi)],$$

and h does not depend on θ. Along a path $\theta(t) = \theta^\star + t\,\delta\theta$ we have

$$\frac{d}{dt}\Big|_{t=0}\,\mathbb{E}_{y\sim p_{\theta(t)}(\cdot|x)}[h(x,y;\xi)] \;=\; \mathbb{E}_{y\sim p^\star(\cdot|x)}\big[h(x,y;\xi)\,s(x,y)^\top\delta\theta\big],$$

where $s(x,y) := \nabla_\theta \log p_\theta(y\mid x)\big|_{\theta=\theta^\star}$. In the linear-softmax model (18),

$$\log p_\theta(y\mid x) = \langle\theta,\Phi(x,y)\rangle - \log\sum_{\tilde y}\exp\big\{\langle\theta,\Phi(x,\tilde y)\rangle\big\},$$

so

$$s(x,y) = \Phi(x,y) - \mathbb{E}_{\tilde y\sim p^\star(\cdot|x)}[\Phi(x,\tilde y)] = \Phi(x,y) - \mu(x).$$

Taking expectation over x and reading off the coefficient of $\delta\theta$ gives the claimed gradient expression $g = \mathbb{E}_x\mathbb{E}_{y\sim p^\star}[h(x,y;\xi)s(x,y)]$. The covariance form follows by expanding $s(x,y) = \Phi(x,y) - \mu(x)$. □

Proposition C.3 (Second-order CE cost and first-order reward gain)**.** *Let $\delta\theta^\dagger = \lambda^{-1}F^{-1}g$ be the optimizer of the local problem (22), with F and g as in (A2) and Lemma C.2. Define the* MARKTUNE*-refined weight $\theta_{\text{MARKTUNE}} = \theta^\star + \delta\theta^\dagger$ and set $\kappa := \lambda^{-1}$. Then, as $\kappa \to 0$,*

$$\mathcal{L}(\theta_{\text{MARKTUNE}}) - \mathcal{L}(\theta^\star) \;=\; \frac{\kappa^2}{2}\, g^\top F^{-1} g \;+\; o(\kappa^2),$$

and

$$\mathcal{R}(\theta_{\text{MARKTUNE}}) - \mathcal{R}(\theta^\star) \;=\; \kappa\, g^\top F^{-1} g \;+\; o(\kappa).$$

Consequently, for a nonzero reward gradient, along these directions in the full parameter space the unnormalized detection statistic experiences a mean shift linear in κ, while the cross entropy deviation from $p^\star$ grows only quadratically in κ.

Proof. By the quadratic expansion (A2),

$$\mathcal{L}(\theta^\star + \delta) - \mathcal{L}(\theta^\star) = \frac{1}{2}\,\delta^\top F\,\delta + o\big(\|\delta\|^2\big).$$

Plugging in $\delta = \delta\theta^\dagger = \lambda^{-1}F^{-1}g$ gives

$$\frac{1}{2}\,(\lambda^{-1}F^{-1}g)^\top F\,(\lambda^{-1}F^{-1}g) = \frac{1}{2\lambda^2}\,g^\top F^{-1}g = \frac{\kappa^2}{2}\,g^\top F^{-1}g,$$

plus $o(\kappa^2)$.

For the reward, the first-order expansion (A3) with $\delta = \delta\theta^\dagger$ gives

$$\mathcal{R}(\theta^\star + \delta) - \mathcal{R}(\theta^\star) = g^\top\delta + o\big(\|\delta\|\big) = \lambda^{-1}\,g^\top F^{-1}g + o(\lambda^{-1}) = \kappa\,g^\top F^{-1}g + o(\kappa).$$

This proves the claim. □

Implications for the improved trade-off. Let $\psi(y;\xi)$ denote the (normalized) GAUSSMARK test statistic computed with the fixed reference model p_θ. Under $\mathbf{H_0}$ (unwatermarked text), with probability taken over the independent Gaussian key, $\mathbb{E}[\psi] = 0$ and the statistic has variance calibrated to yield a level-α test (Proposition 3.3 in the main text).

The same local argument applies to $\mathcal{R}_\psi(\theta) = \mathbb{E}_x\mathbb{E}_{p_\theta}[\psi(y,\xi_\sigma \mid x)]$, provided the corresponding differentiability and integrability conditions hold. Its gradient is $g_\psi = \mathbb{E}_x\,\mathrm{Cov}_{p^\star}(\psi,\Phi)$, and the update is $\kappa F^{-1}g_\psi$. Thus the normalized mean gains $\kappa g_\psi^\top F^{-1}g_\psi + o(\kappa)$, with cross-entropy cost $\frac{1}{2}\kappa^2 g_\psi^\top F^{-1}g_\psi + o(\kappa^2)$. This is a distinct update from the one derived for h unless their gradients coincide up to scale. Smoothness alone does not imply constant variance or increasing tail probability: for a direction v, the directional derivative of detection power is

$$\frac{d}{dt}\Pr_{p_{\theta^\star+tv}}(\psi \ge \tau_\alpha)\bigg|_{t=0} = \mathbb{E}_x\,\mathrm{Cov}_{p^\star}(\mathbf{1}\{\psi \ge \tau_\alpha\},\Phi)^\top\, v,$$

whose sign is not determined by the mean-reward gain. The population power theorem in Appendix B.4 and the empirical detection measurements therefore complement this local quality-reward analysis.

The mean-shift intuition has a precise conditional interpretation. Let $C_\psi = g_\psi^\top F^{-1}g_\psi > 0$, and suppose, in addition, that under $p_{\theta^\star+\kappa F^{-1}g_\psi}$ the score distribution is a local location shift: its CDF satisfies $F_\kappa(t) = F_0(t - \kappa C_\psi) + o(\kappa)$ at $t = \tau_\alpha$, with baseline density $f_0(\tau_\alpha) > 0$. Then the normalized reward increment corresponds to a mean shift of order κC_ψ, and detection power increases by $\kappa C_\psi f_0(\tau_\alpha) + o(\kappa)$ at the fixed null threshold, while the cross-entropy cost remains second order. Under an exact location shift the variance is unchanged; this is an additional distributional condition, not a consequence of smoothness alone. This recovers the first-order power versus second-order quality interpretation when the score law has the required form.

Soundness of the stylized model and local assumptions. The derivation assumes that the local Fisher geometry at $\theta^\star$ is well-behaved (so that F is positive definite on the relevant subspace) and that the linear-softmax model (18) adequately captures the first-order behavior of the true language model. In highly over-parameterized language models with wide, flat basins, such local expressivity is often a reasonable approximation [4]. From a geometric viewpoint, $\delta\theta^\dagger$ can be seen as the Riesz representer of the linear functional g under the Fisher metric F; any parameterization that is locally surjective in the neighborhood of $\theta^\star$ admits such a direction.

Table 3: Hyperparameter setting for MARKTUNE on Qwen3-4B-Base.

Model	Qwen3-4B-Base on OpenWebText
Training steps	200
GRPO inner steps	3
Regularization term	Cross-entropy (CE) regularization
Regularization coefficient λ	0.01
Optimizer	AdamW
maximal learning rate	5e-6
Learning rate schedule	Cosine decay with 20 steps linear warmup
Learning rate decay ratio (training steps)	40%
Momentum (β_1, β_2)	(0.9, 0.97)
Group size	8
Prompt batch size	64
Mini batch size	64
Max sampling sequence length	256 tokens
Sampling temperature	0.7
CE batch size	128
CE sequence length	512 tokens

Table 4: Hyperparameter setting for MARKTUNE on Llama2-7B.

Model	Llama2-7B on OpenWebText
Training steps	300
GRPO inner steps	3
Regularization term	Cross-entropy (CE) regularization
Regularization coefficient λ	0.005
Optimizer	AdamW
maximal learning rate	4e-6
Learning rate schedule	Cosine decay with 30 steps linear warmup
Learning rate decay ratio (training steps)	40%
Momentum (β_1, β_2)	(0.9, 0.97)
Group size	8
Prompt batch size	64
Mini batch size	64
Max sampling sequence length	256 tokens
Sampling temperature	1.0
CE batch size	128
CE sequence length	512 tokens

D Implementation Details

In this section, we provide details on the implementation of metrics, baselines, our methods, and the fine-tuning attack. We run the experiments on 4 NVIDIA H200 GPUs.

D.1 Evaluation metrics

For the MAUVE metric, we select samples from the **RealNewsLike** split of **C4** with text longer than 400 tokens, using the first 200 tokens as the model prompt and the next 200 tokens as the human-generated reference. For the **ELI5** dataset, we select samples whose answers exceed 200 tokens and truncate each answer to 200 tokens to serve as the human-generated reference.

In Figures 3 and 6, dashed lines correspond to a bootstrap estimate (500 resamples) of the mean TPR@FPR=1%, and the shaded regions correspond to the 90% confidence interval on the mean estimate.

D.2 Watermarking methods

Baselines. We use the inference-time KGW watermark [42], which biases the output logits toward a PRF-determined "green list" based on the token prefix. We set the token context length to $k = 4$, the green-list fraction to $\gamma = 0.5$, the bias strength to $\delta = 1$ for KGW-1, and $\delta = 2$ for KGW-2. For Figure 3, we sweep the bias strength δ over the set $\{0.6, 0.7, 0.8, 0.9, 1.0, 1.2, 1.5\}$. For the Gumbel-max watermark [1] and the SynthID watermark [13], we also use a 4-token prefix for fair comparison. For SynthID, we set the number of tournaments to 30, following the recommended default setting. For KGW-D [27], we distill the logit-based KGW with $\gamma = 0.25$ and $\delta = 1$ to preserve generation quality. Note that we distill the 1-token-prefix KGW variant, since watermarks with longer token prefixes are extremely difficult for the model to learn. The KGW, Gumbel-max, and SynthID baselines are implemented on top of the open-source MarkLLM toolkit [54], which provides reference implementations and unified inference and detection interfaces for the inference-time schemes considered here.

For GAUSSMARK [4], given a vectorized target parameter θ_{wm} with dimension d_r, we first compute its RMS norm $\|\theta_{\mathrm{wm}}\|_{\mathrm{RMS}} = \|\theta_{\mathrm{wm}}\|_{\mathrm{F}}/\sqrt{d_r}$. The watermark key is then sampled as $\xi_\sigma \sim \mathcal{N}(0,\, \sigma^2\|\theta_{\mathrm{wm}}\|^2_{\mathrm{RMS}}\, \mathbb{I}_{d_r})$. This gives $\mathbb{E}\,\|\xi_\sigma\|^2_{\mathrm{F}} \approx \sigma^2\|\theta_{\mathrm{wm}}\|^2_{\mathrm{RMS}} d_r$, and thus $\mathbb{E}[\|\xi_\sigma\|_{\mathrm{F}}/\|\theta_{\mathrm{wm}}\|_{\mathrm{F}}] \approx \sigma$. This normalization ensures that the relative perturbation magnitude is directly controlled by σ, making watermark strength easy to tune across different parameter scales and model architectures. We perform a grid search over candidate target parameters and select the up-projection MLP matrix in layer 28 for Qwen3-4B-Base with $\sigma = 1.0$, and the up-projection MLP matrix in layer 30 for Llama2-7B with $\sigma = 1.2$. For Figure 3, we sweep the perturbation strength σ over the set $\{0.6, 0.8, 1.0, 1.1, 1.2, 1.5, 1.8\}$ for Qwen3-4B-Base and $\{0.8, 1.0, 1.2, 1.3, 1.4, 1.6, 2.0\}$ for Llama2-7B.

Our method. For MARKTUNE, we inherit the target parameter from GAUSSMARK and use $\sigma = 0.6$ for Qwen3-4B-Base and $\sigma = 0.8$ for Llama2-7B to apply a soft activation. For the KL regularization, we use the Qwen3-8B and Llama2-13B as the oracle models, respectively. The GRPO-related hyperparameters are listed in Table 3 and Table 4. For Figure 3, we sweep the number of training steps over the set $\{50, 100, 150, 200, 250, 300, 500\}$ for Qwen3-4B-Base and $\{50, 100, 200, 300, 400, 500, 600\}$ for Llama2-7B.

D.3 Fine-tuning attack

We fine-tune models on **OpenWebText** for 1500 steps using the AdamW optimizer with a maximal learning rate of 1×10^{-5} and a cosine decay schedule with 300 warmup steps. We use a sequence length of 512 and a batch size of 64. For LoRA, we set the rank to 8 and the alpha to 16. LoRA is applied to all MLP projection layers as well as to the unembedding layer.

D.4 Computation time benchmark

We benchmark wall-clock times at inference time across five watermarking schemes: GAUSSMARK, MARKTUNE (ours), KGW, Gumbel-max, and SynthID. Prompts are taken from the `realnewslike` split of **C4** [58]: we use the first 100 prompts truncated to a 200-token prefix and generate 200 new tokens with batch size 1 on a single H200 GPU. Every method is run through the standard HuggingFace `model.generate` pipeline. The inference-time schemes are wrapped via MarkLLM's `LogitsProcessor` interface, while GAUSSMARK and MARKTUNE alter only model weights and require no per-step intervention. The three hash-window schemes share an identical context length of 4 tokens. Times are bracketed by `torch.cuda.synchronize()` and the first two prompts of each method are discarded as warm-up. Reported values are mean $\pm$ std over the remaining $n = 100$ prompts. Per-prompt numbers are listed in Appendix D.6 (Table 5).

D.5 Detection access and memory

Detection requires the candidate text, the secret watermark key, and the fixed clean reference model. It consists of one forward pass through the clean reference and one backward pass to compute the gradient with respect to the selected watermarked parameter block. With the Qwen3-4B-Base reference, a batch size of 64, and 200-token suffixes, peak allocated GPU memory is 107.78 GiB,

which fits on a single H200 GPU. Generation uses the released watermarked model with standard decoding and incurs no watermark-specific inference overhead.

D.6 Computational Cost

Training cost. The full MARKTUNE pipeline adds a one-time, offline training stage on top of GAUSSMARK; afterwards generation and detection are identical to GAUSSMARK. Training uses 4 NVIDIA H200 GPUs with ≈8 hours for Qwen3-4B-Base (200 gradient steps) and ≈13.5 hours for Llama2-7B (300 steps). This is modest relative to either model's pre-training budget, and substantially cheaper than typical RLVR fine-tuning for reasoning, since MARKTUNE rollouts terminate at 256 tokens (Table 3) rather than the multi-thousand-token chains-of-thought required for reasoning RL.

Inference-time cost. Table 5 reports per-prompt wall-clock generation and detection times for the five watermarking schemes; the corresponding bar plot is given in Figure 4 of the main text and the benchmark protocol is detailed in Appendix D.4. Across both models, GAUSSMARK and MARKTUNE generate text within 0.7–1.5% of each other and are 9–16% faster than KGW, Gumbel-max, and SynthID, since each inference-time decoding step has to materialize a green-list mask, a Gumbel pseudo-random table, or a hash-tournament sampler. On detection, GAUSSMARK and MARKTUNE—which perform a single forward+backward over the response and project the gradient onto the secret noise direction ξ_σ—are faster in these measurements than the inference-time baselines, with 18–27 ms versus 35–176 ms.

Table 5: Per-prompt watermarked generation and detection times on the first $n = 100$ **C4-RealNewsLike** prompts (200 prompt tokens, 200 generated tokens, batch size 1, single H200, bf16, single-prompt HuggingFace `model.generate`). Means ± standard deviations are computed after a 2-prompt warm-up.

Method	Generation (ms)	Detection (ms)
Qwen3-4B-Base		
GAUSSMARK	3371.3 ± 8.1	26.2 ± 0.2
MARKTUNE (ours)	3422.8 ± 5.7	27.0 ± 0.2
KGW	3714.6 ± 11.0	69.7 ± 0.2
Gumbel-max	3716.9 ± 5.4	175.7 ± 0.9
SynthID	3971.9 ± 10.2	35.5 ± 0.2
Llama2-7B		
GAUSSMARK	2421.8 ± 6.8	18.1 ± 0.0
MARKTUNE (ours)	2404.9 ± 5.3	18.1 ± 0.0
KGW	2654.3 ± 3.1	62.3 ± 0.1
Gumbel-max	2555.0 ± 4.0	42.4 ± 0.2
SynthID	2882.9 ± 7.2	35.0 ± 0.1

E Ablation Study

We isolate the roles of Gaussian initialization, the watermark reward, and quality regularization, and examine the sensitivity to GRPO hyperparameters and the selected parameter block. Unless otherwise specified, these experiments use Qwen3-4B-Base and C4-RealNewsLike, with TPR measured at 1% FPR.

E.1 Gaussian initialization

To isolate the role of Gaussian soft activation, we compare MARKTUNE with direct GRPO from the clean base model under two CE coefficients. All trained models use the same 200-step protocol and differ only in initialization and the CE coefficient. Table 6 shows that direct GRPO with $\lambda = 0.01$ achieves high detectability but produces repetitive, low-quality text, consistent with reward hacking. Increasing the coefficient to 0.1 largely restores generation quality, at the cost of lower detection power.

Table 7 evaluates the same models under subsequent LoRA fine-tuning. Direct GRPO loses detection power more rapidly: after 1500 steps, TPR falls to 0.496 and 0.384 for the two CE coefficients,

Table 6: Gaussian initialization ablation on Qwen3-4B-Base/C4-RealNewsLike. TPR is measured at 1% FPR. PPL subscripts denote standard deviations.

Method	AUC	TPR	PPL	Seq-rep-3	MAUVE
Clean base model	0.500	0.010	$4.92_{1.34}$	0.04	0.98
MARKTUNE(GAUSSMARKinit., CE $= 0.01$)	0.995	0.964	$5.04_{1.37}$	0.03	0.97
Direct GRPO (CE $= 0.01$)	0.994	0.952	$6.21_{1.79}$	0.16	0.81
Direct GRPO (CE $= 0.1$)	0.984	0.812	$5.01_{1.33}$	0.04	0.97

compared with 0.730 for MARKTUNE. Together, these controls support Gaussian initialization as a useful starting point for quality-preserving watermark optimization and robustness to continued model adaptation. A computable detection reward alone does not supply this initialization: KGW and Gumbel-max inject their signals at decoding time rather than through a soft perturbation of the model weights.

Table 7: Detection robustness of MARKTUNE and direct GRPO under LoRA fine-tuning. Entries are TPR at 1% FPR.

Fine-tuning steps	MARKTUNE	Direct GRPO (CE $= 0.01$)	Direct GRPO (CE $= 0.1$)
0	0.964	0.952	0.812
300	0.892	0.814	0.698
900	0.784	0.668	0.522
1500	0.730	0.496	0.384

E.2 Regularization term and coefficient

Table 8 highlights the effect of the regularization term. Without regularization, the model aggressively amplifies the watermark signal but at a cost to generation quality. Introducing cross-entropy or KL regularization restores perplexity and diversity while maintaining strong detectability, with cross-entropy with $\lambda = 0.01$ offering the most balanced performance and thus serving as our default choice in Sections 4.2 and 4.3. This shows that MARKTUNE's gains over GAUSSMARK do not come from simply increasing watermark strength, but from guiding weight updates in a watermark-sensitive direction while staying close to the underlying high-quality text generation distribution.

Table 8: Ablation study on regularization hyperparameters. "Base" denotes the watermarked model trained using MARKTUNE without a regularization term. Configuration highlighted in **bold** indicates Pareto-optimal performance.

Method	**AUC↑**	**TPR↑**	**PPL↓**	**Seq-rep-3↓**	**MAUVE↑**
Base	0.999	0.992	$5.88_{1.55}$	0.06	0.89
+ CE-0.005	0.997	0.976	$5.35_{1.43}$	0.05	0.94
+ CE-0.01	**0.995**	**0.964**	$\mathbf{5.04_{1.37}}$	**0.03**	**0.97**
+ CE-0.05	0.989	0.902	$4.97_{1.31}$	0.03	0.97
+ KL-0.001	0.997	0.986	$5.69_{1.56}$	0.06	0.93
+ KL-0.005	0.993	0.930	$5.36_{1.41}$	0.06	0.95
+ KL-0.01	0.986	0.854	$5.13_{1.32}$	0.04	0.97

E.3 Choice of quality anchor

A central design choice in MARKTUNE is to regularize toward an external high-quality distribution (cross-entropy on human text or KL to a stronger oracle), rather than toward the base model as in standard RLHF/RLVR. The motivation is that, for model-embedded watermarking, increasing the watermark detectability inherently moves the trained model away from the base model; if the regularizer also penalizes deviation from the base model, the reward and the regularizer become directly antagonistic and this penalty can limit improvements over the existing GAUSSMARK frontier. Anchoring instead to an external high-quality reference decouples the two objectives, making room for a strictly better trade-off. The fixed clean reference defines the statistic, whereas the external

anchor defines the language-model regularization loss. Neither human-text cross-entropy nor base-model KL is a complete quality measure; an external anchor need not dominate the base model on every task.

To validate this, we replace $\mathcal{L}_{\text{reg}}$ with the standard RL-style anchor $\mathrm{D}_{\mathrm{KL}}\big(p_{\theta(\xi_\sigma)}(\cdot \mid x) \,\|\, p_{\theta_{\text{ref}}}(\cdot \mid x)\big)$ (KL to the frozen base model), keeping all other hyperparameters identical to the Qwen3-4B-Base configuration in Table 3. We track two checkpoints chosen to match the quality and the detectability of MARKTUNE (CE-human), respectively. Results on **C4-RealNewsLike** are reported in Table 9. Anchoring to the base model yields essentially no improvement over GAUSSMARK at matched quality, and only modest detectability gains at the cost of perceptible quality degradation. In contrast, anchoring to an external high-quality distribution produces a strictly better point on the trade-off frontier.

Table 9: Ablation on the regularization anchor (Qwen3-4B-Base, **C4-RealNewsLike**). KL-to-base-model regularization fails to improve over GAUSSMARK at matched quality and only matches MARKTUNE (CE-human) by sacrificing generation quality.

Method	**AUC↑**	**TPR↑**	**PPL↓**	**Seq-rep-3↓**	**MAUVE↑**
GAUSSMARK	0.983	0.788	$5.16_{1.54}$	0.04	0.98
MARKTUNE + KL-to-base (quality-matched)	0.986	0.796	$5.09_{1.52}$	0.03	0.98
MARKTUNE + KL-to-base (detectability-matched)	0.993	0.940	$5.46_{1.60}$	0.03	0.94
MARKTUNE + CE-human (default)	**0.995**	**0.964**	$\mathbf{5.04_{1.37}}$	**0.03**	**0.97**

Role of the regularization corpus. OpenWebText supplies both the RL prompts and the text for the CE regularizer. We use it as a regularization distribution; we neither search for nor curate a corpus that improves the pretrained base model. The anchor need not be globally better than the base model. Its role is to constrain quality loss while the watermark reward selects useful directions for increasing detectability. The local Fisher-geometry analysis in Appendix C explains how these objectives can be compatible even near a strong quality reference.

Quality-only fine-tuning. To measure the opportunity cost of watermarking under the same training procedure, we fine-tune the clean, unwatermarked base model on the regularization corpus using CE alone for 200 steps. Table 10 compares this control with the clean base model and MARKTUNE. CE-only fine-tuning yields no clear overall quality improvement under PPL, Seq-rep-3, and MAUVE. Thus the observed watermarking gains cannot be explained simply by access to a corpus that improves the base model.

Table 10: Generation quality of the clean base model, matched quality-only fine-tuning, and MARKTUNE on Qwen3-4B-Base/C4-RealNewsLike. PPL subscripts denote standard deviations.

Method	PPL	Seq-rep-3	MAUVE
No Watermark (base model)	$4.92_{1.34}$	0.04	0.98
No Watermark (CE-only FT)	$4.97_{1.37}$	0.03	0.98
MARKTUNE	$5.04_{1.37}$	0.03	0.97

These generation-quality measures are distinct from KL divergence to the base model in Proposition 3.1, which characterizes GAUSSMARK's distributional distortion. The regularizer guides optimization; oracle PPL, repetition, MAUVE, and downstream task performance evaluate its effect on utility.

E.4 Removing the watermark reward

We further isolate the role of the watermark reward by ablating it from the training objective entirely, i.e., dropping the first term of (3) so that the objective reduces to minimizing $\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{ce}}$. Concretely, we initialize a strongly perturbed GAUSSMARK model with $\sigma = 1.6$ on Qwen3-4B-Base so that its starting detectability roughly matches that of MARKTUNE, and then perform full-parameter fine-tuning on **OpenWebText** using cross-entropy alone, reusing the optimizer and learning-rate schedule of Table 3. This is the natural model-embedded analogue of the LoRA fine-tuning attack

in Section 4.3 but applied without any LoRA constraint. As shown in Table 11, the quality-driven fine-tuning steadily moves the model back toward the base distribution and erodes the watermark. After 1000 steps, the resulting model is Pareto-dominated by MARKTUNE on both axes (TPR 0.726 vs. 0.964, PPL 5.37 vs. 5.04). This confirms that $\mathcal{R}_{\text{wm}}$ is necessary: without it, cross-entropy optimization treats the perturbation as quality damage to be undone, while MARKTUNE uses $\mathcal{R}_{\text{wm}}$ to steer updates along watermark-sensitive directions that preserve, rather than dissipate, the signal.

Table 11: Removing the watermark reward $\mathcal{R}_{\text{wm}}$ during fine-tuning (Qwen3-4B-Base, **C4-RealNewsLike**). Starting from GAUSSMARK with $\sigma = 1.6$, cross-entropy-only fine-tuning progressively erases the watermark; MARKTUNE Pareto-dominates the result on both detectability and quality.

Method	AUC↑	TPR↑	PPL↓	Seq-rep-3↓	MAUVE↑
No Watermark	0.500	0.010	$4.92_{1.34}$	0.04	0.98
GAUSSMARK ($\sigma = 1.6$, init)	0.995	0.970	$5.79_{1.68}$	0.05	0.93
w/o $\mathcal{R}_{\text{wm}}$ (CE only, 200 steps)	0.990	0.888	$5.66_{1.64}$	0.05	0.93
w/o $\mathcal{R}_{\text{wm}}$ (CE only, 1000 steps)	0.973	0.726	$5.37_{1.49}$	0.03	0.95
MARKTUNE (default)	**0.995**	**0.964**	$\mathbf{5.04_{1.37}}$	**0.03**	**0.97**

E.5 GRPO sampling hyperparameters

As for the impact of GRPO sampling hyperparameters on MARKTUNE, Table 12 indicates that the method is generally insensitive to the choice of group size. In practice, a group size of 8 provides the best balance between performance and computational cost. The sampling temperature should follow the model-specific settings commonly used for high-quality text generation, as deviations from these values either reduce output diversity or increase perplexity.

Table 12: Ablation study on GRPO sampling hyperparameters group size ("**G**") and temperature ("**T**"). Configuration highlighted in **bold** indicates Pareto-optimal performance.

G	T	AUC↑	TPR↑	PPL↓	Seq-rep-3↓	MAUVE↑
8	**0.7**	**0.995**	**0.964**	$\mathbf{5.04_{1.37}}$	**0.03**	**0.97**
4	0.7	0.992	0.900	$5.07_{1.38}$	0.03	0.98
16	0.7	0.996	0.968	$5.11_{1.42}$	0.03	0.97
8	0.5	0.983	0.820	$4.98_{1.28}$	0.06	0.92
8	1.0	0.992	0.940	$6.21_{2.26}$	0.03	0.94

E.6 Alternative watermark rewards

Although our analysis and main experiments instantiate MARKTUNE with the GAUSSMARK test statistic as the watermark reward $\mathcal{R}_{\text{wm}}$, the framework itself does not depend on this specific choice. As stated in Remark 3.2, a broad class of computable detection statistics can in principle serve as the reward. To probe this generality, we replace the reward in (3) with the KGW z-score [42] ($\gamma = 0.5$, $\delta = 0.6$, prefix length $k = 1$) while reusing the Qwen3-4B-Base optimizer and regularization hyperparameters of Table 3. Table 13 reports the resulting model on **C4-RealNewsLike** alongside two pure-KGW baselines that bracket its detectability. With the KGW reward, MARKTUNE improves over the stronger KGW configuration ($\delta = 0.8$) in both detectability (AUC, TPR) and perplexity, indicating that the framework's quality-preserving optimization geometry is not specific to the GaussMark statistic. Note that the GaussMark-rewarded MARKTUNE still outperforms its KGW-rewarded counterpart, which we attribute to the cleaner statistical and theoretical inheritance available for GAUSSMARK.

E.7 Choice of watermarked parameters

MARKTUNE inherits the layer and parameter matrix selected by the GAUSSMARK search. Because it optimizes the same detector statistic, this provides a natural candidate without exhaustively retraining MARKTUNE for every parameter block. We examine nearby late-layer MLP projections

Table 13: MARKTUNE with the KGW z-score as reward (Qwen3-4B-Base, **C4-RealNewsLike**). MARKTUNE (KGW) improves on KGW ($\delta = 0.8$) in both detectability and quality.

Method	AUC↑	TPR↑	PPL↓	Seq-rep-3↓	MAUVE↑
No Watermark	0.500	0.010	$4.92_{1.34}$	0.04	0.98
KGW ($\gamma = 0.5$, $\delta = 0.6$)	0.935	0.513	$4.95_{1.44}$	0.04	0.98
KGW ($\gamma = 0.5$, $\delta = 0.8$)	0.987	0.861	$5.14_{1.50}$	0.05	0.98
MARKTUNE (KGW reward)	**0.991**	**0.906**	$\mathbf{5.09_{1.49}}$	0.06	**0.98**
MARKTUNE (GAUSSMARK reward, default)	0.995	0.964	$5.04_{1.37}$	0.03	0.97

on both model families in Table 14. The evaluated up-projection modules have comparable quality–detectability trade-offs, whereas the down-projection alternatives have weaker detection and no clear quality advantage. These comparisons support the late-layer up-projection choice within this candidate set; they do not establish global optimality.

Table 14: Watermarking performance for nearby parameter matrices. TPR is measured at 1% FPR; PPL subscripts denote standard deviations.

Target	TPR	PPL	Seq-rep-3
Qwen3-4B-Base			
L26 `mlp.up_proj`	0.822	$5.28_{1.81}$	0.04
L28 `mlp.up_proj` (default)	0.788	$5.16_{1.54}$	0.04
L30 `mlp.up_proj`	0.808	$5.47_{1.88}$	0.05
L28 `mlp.down_proj`	0.610	$5.66_{1.86}$	0.05
Llama2-7B			
L28 `mlp.up_proj`	0.650	$6.70_{1.97}$	0.01
L30 `mlp.up_proj` (default)	0.712	$6.78_{1.91}$	0.01
L30 `mlp.down_proj`	0.586	$7.12_{2.01}$	0.01

The search evaluates GAUSSMARK on 500 prompts per candidate, without retraining MARKTUNE for each candidate. The complete search takes 66 minutes for Qwen3-4B-Base and 41 minutes for Llama2-7B on a single H200 GPU.

F Supplemental Results

This section extends the main evaluation with low-FPR detection, additional baselines, decoder and length sensitivity, and text- and model-level attacks.

F.1 Detection at ultra low false-positive rates

AUC summarizes ranking quality across thresholds, but neither AUC nor TPR at 1% FPR characterizes the extreme left tail of the ROC curve. We therefore also report mean and median $\log_{10} p$ under the alternative and TPR at FPR@10^{-4}. For GAUSSMARK and MARKTUNE, the one-sided p-value in Eq. (8) is uniform under the key-independent null by Appendix B.2.

Table 15: Low-FPR detection and generation quality on Qwen3-4B-Base/C4-RealNewsLike. Brackets give 95% bootstrap confidence intervals over 500 prompts. PPL subscripts denote standard deviations.

Method	Mean $\log_{10} p$	Median $\log_{10} p$	TPR at FPR 10^{-4}	PPL	Seq-rep-3	MAUVE
KGW-1	-6.106	-5.916	0.806 [0.770,0.840]	$5.51_{1.56}$	0.04	0.98
KGW-D	-3.320	-3.062	0.302 [0.268,0.336]	$5.54_{1.53}$	0.04	0.98
GAUSSMARK	-4.265	-3.923	0.484 [0.440,0.528]	$5.16_{1.54}$	0.04	0.98
MARKTUNE	-6.014	-5.871	0.800 [0.764,0.836]	$5.04_{1.37}$	0.03	0.97

MARKTUNE improves on GAUSSMARK and KGW-D in both log-p-value summaries and low-FPR detection power. At FPR 10^{-4}, its TPR is 0.800, compared with 0.484 for GAUSSMARK and 0.302

for KGW-D. MARKTUNE and KGW-1 perform similarly in this comparison; the reported results do not establish a statistically significant quality advantage for either method. KGW-1 remains an inference-time reference that requires control of decoding, whereas MARKTUNE supports open-weight deployment.

F.2 Multi-key deployment

In a realistic deployment, a provider may release multiple watermarked copies of the same base model, each with a different secret key. The relevant statistical question is whether keys collide: when the detector verifies one copy with the wrong key, does the test still trigger? In GAUSSMARK, each key ξ_σ is sampled as an isotropic Gaussian vector in a high-dimensional parameter subspace. Two independent draws are therefore nearly orthogonal with overwhelming probability: their cosine similarity scales as $\mathcal{O}(1/\sqrt{d_r})$, which is negligible whenever the watermark subspace dimension d_r is large (the configurations used in this paper have $d_r \geq 10^7$). MARKTUNE inherits this property exactly because the key sampling distribution is unchanged.

To validate this empirically, we sample ten independent keys $\{\xi_i\}_{i=1}^{10}$ on Qwen3-4B-Base and run detection on text generated by the MARKTUNE-trained model that was watermarked with a different key $\xi^\star$. Table 16 reports the resulting TPR@1%FPR on **C4-RealNewsLike**. Every mismatched key produces a TPR within Monte Carlo fluctuation of the nominal level $\alpha = 0.01$, confirming that cross-key false positives in a multi-key deployment are negligible at this scale; combined with the fact that each detector uses only its own key and the reference model, this implies that the matched-key detection power for any given key is unaffected by the simultaneous presence of additional independently sampled keys.

Table 16: Cross-key detection on Qwen3-4B-Base/**C4-RealNewsLike**: each cell reports TPR@1%FPR when the detector uses a key ξ_i different from the key $\xi^\star$ used to watermark the generator. All entries lie at or near the nominal $\alpha = 0.01$ level.

Mismatched key	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6	ξ_7	ξ_8	ξ_9	ξ_{10}
TPR@1%FPR	0.012	0.008	0.014	0.014	0.010	0.012	0.012	0.014	0.010	0.012

F.3 Additional model-embedded baseline

We compare against the training-free model-embedded watermark of Christ et al. [9]. Because Qwen3-4B-Base has no output-layer bias, we introduce a vocabulary-sized output bias parameter to apply the method, following its original design as closely as possible. Table 17 evaluates three perturbation strengths alongside the model-embedded baselines.

Table 17: Model-embedded watermarking methods on Qwen3-4B-Base/C4-RealNewsLike. TPR is measured at 1% FPR; PPL subscripts denote standard deviations.

Method	AUC	TPR	PPL	Seq-rep-3
Christ ($\epsilon = 0.1$)	0.672	0.026	$5.17_{1.60}$	0.05
Christ ($\epsilon = 0.3$)	0.930	0.308	$5.25_{1.57}$	0.05
Christ ($\epsilon = 0.6$)	0.993	0.884	$5.60_{1.74}$	0.07
KGW-D	0.981	0.762	$5.54_{1.53}$	0.04
GAUSSMARK	0.983	0.788	$5.16_{1.54}$	0.04
MARKTUNE	0.995	0.964	$5.04_{1.37}$	0.03

Increasing the output-bias perturbation improves detectability but raises perplexity and repetition. Among the evaluated model-embedded methods, MARKTUNE achieves the highest AUC and TPR with the lowest PPL and Seq-rep-3.

F.4 Decoding strategy and output length

MARKTUNE does not require a watermark-specific decoding rule, but changing the decoder changes the output distribution and can affect both detectability and quality. Table 18 compares multinomial

sampling at three temperatures, nucleus sampling, and top-k sampling under the same evaluation protocol. Detection remains strong across all five settings. Higher temperature increases diversity at the cost of perplexity.

Table 18: MARKTUNE under standard decoding strategies on Qwen3-4B-Base/C4-RealNewsLike. TPR is measured at 1% FPR; PPL subscripts denote standard deviations.

Decoding strategy	AUC	TPR	PPL	Seq-rep-3	MAUVE
Multinomial ($T = 0.7$) (default)	0.995	0.964	$5.04_{1.37}$	0.03	0.97
Multinomial ($T = 0.5$)	0.994	0.942	$3.91_{1.00}$	0.07	0.94
Multinomial ($T = 1.0$)	0.996	0.982	$12.36_{4.66}$	0.01	0.91
Nucleus ($T = 0.7,\ p = 0.9$)	0.993	0.952	$4.20_{1.12}$	0.04	0.96
Top-k ($T = 0.7,\ k = 50$)	0.993	0.944	$4.98_{1.37}$	0.03	0.97

We next vary the generation length while holding the model, prompts, key, and decoding strategy fixed. Table 19 shows that detection improves with more generated tokens and then saturates: TPR rises from 0.414 at 50 tokens to 0.964 at 200 tokens and remains near 0.99 at 400–500 tokens. Longer outputs provide more evidence for detection, while also exhibiting greater repetition.

Table 19: Detection and generation quality across output lengths. TPR is measured at 1% FPR; PPL subscripts denote standard deviations.

Generated tokens	AUC	TPR	PPL	Seq-rep-3
50	0.907	0.414	$6.50_{2.75}$	0.01
100	0.971	0.736	$5.97_{1.79}$	0.02
200 (default)	0.995	0.964	$5.04_{1.37}$	0.03
300	0.997	0.986	$4.60_{1.26}$	0.05
400	0.999	0.992	$4.07_{1.16}$	0.06
500	1.000	0.990	$3.75_{1.03}$	0.08

F.5 Robustness to text editing

Token-edit strength. We evaluate random insertion, word deletion, and synonym substitution at seven edit ratios from 0.02 to 0.5. Table 20 reports TPR at 1% FPR on Qwen3-4B-Base/C4-RealNewsLike. Detection decreases with the edit ratio for both methods, and MARKTUNE retains higher TPR than GAUSSMARK throughout the evaluated grid. At large deletion or substitution ratios, both signals become weak.

Table 20: TPR at 1% FPR under seven token-edit ratios on Qwen3-4B-Base/C4-RealNewsLike.

Method	Attack	0.02	0.05	0.1	0.2	0.3	0.4	0.5
GAUSSMARK	Random insertion	0.842	0.802	0.738	0.518	0.358	0.220	0.126
GAUSSMARK	Word deletion	0.800	0.722	0.626	0.366	0.138	0.060	0.048
GAUSSMARK	Synonym substitution	0.800	0.748	0.612	0.384	0.264	0.138	0.100
MARKTUNE	Random insertion	0.914	0.900	0.858	0.734	0.604	0.434	0.310
MARKTUNE	Word deletion	0.896	0.850	0.766	0.484	0.226	0.100	0.068
MARKTUNE	Synonym substitution	0.912	0.870	0.766	0.552	0.292	0.168	0.110

Semantic-preserving attacks. Random edits control the amount of surface modification, but can substantially degrade fluency and meaning. We therefore also evaluate Dipper paraphrasing and roundtrip translation, which better preserve the utility of the text. Table 21 reports their mean realized token edit ratios, measured by normalized token-level Levenshtein distance.

The realized edit ratios are not identical across methods. In particular, Dipper-2 changes MARKTUNE outputs less than KGW-2 and KGW-D outputs (0.788 versus 0.871 and 0.879). Relative to GAUSSMARK, MARKTUNE has a lower edit ratio under Dipper-1 but higher ratios under Dipper-2 and translation. The results therefore characterize robustness to these rewriting pipelines; they do not isolate robustness at matched realized edit counts or establish a general tendency for MARKTUNE text to resist modification.

Table 21: Mean realized token edit ratio under Dipper paraphrasing and roundtrip translation.

Method	Dipper-1	Dipper-2	Translate
KGW-1	0.470	0.856	0.488
KGW-2	0.501	0.871	0.521
Gumbel-max	0.462	0.868	0.443
SynthID	0.456	0.761	0.390
KGW-D	0.482	0.879	0.507
GAUSSMARK	0.450	0.761	0.409
MARKTUNE	0.437	0.788	0.427

Global key and attack quality. GAUSSMARK and MARKTUNE use a single Gaussian key in weight space and a full-sequence gradient-alignment statistic. They have no prefix hash window, so a prefix-window sweep and its associated security–robustness trade-off do not apply. Table 24 supplements the main robustness results with post-attack perplexity. Deletion and substitution substantially degrade quality, whereas paraphrasing preserves lower perplexity. These quality measures complement the edit-strength sweep when assessing attacks intended to preserve useful text.

F.6 Quantization, pruning, and model merging

We complement LoRA fine-tuning and logit-based distillation with three open-weight model modifications, following the evaluation protocol of Gloaguen et al. [24]: INT8 quantization, WANDA pruning at $\rho = 0.2$, and SLERP merging at $t = 0.5$. Table 22 reports the detection and generation quality of MARKTUNE before and after each attack.

Table 22: MARKTUNE under open-weight attacks. TPR is measured at 1% FPR; PPL subscripts denote standard deviations.

Attack	AUC	TPR	PPL
No attack	0.995	0.964	$5.04_{1.37}$
Quantization (INT8)	0.981	0.754	$5.27_{1.46}$
Pruning (WANDA, $\rho = 0.2$)	0.988	0.838	$5.49_{1.50}$
Model merging (SLERP, $t = 0.5$)	0.949	0.584	$5.18_{1.42}$

All three attacks reduce detection power while preserving generation quality close to the unmodified model. MARKTUNE retains nontrivial detectability, with TPR 0.754 after quantization, 0.838 after pruning, and 0.584 after merging. These results support robustness within the evaluated settings. Full-parameter fine-tuning of the final MARKTUNE model and cross-architecture distillation remain outside this evaluation.

F.7 Robustness to logit-based distillation

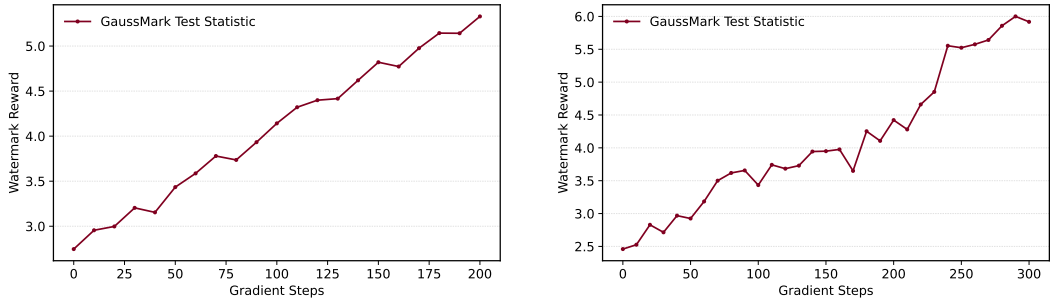
We evaluate whether the watermark embedded by MARKTUNE survives logit-based knowledge distillation from a watermarked teacher [27]. Following the protocol of Gu et al. [27], we instantiate teacher and student as the same architecture, Llama2-7B, with the teacher initialized from a MARKTUNE-trained checkpoint and the student initialized from a fresh copy of the unwatermarked Llama2-7B; the student is then trained on **OpenWebText** for 5000 steps with AdamW, batch size 64, sequence length 512, peak learning rate 1×10^{-5}, and a cosine schedule with linear warmup. Detection is run on the student outputs using the original watermark key. Table 23 shows that the distilled same-size student retains the bulk of the watermark signal (TPR $0.982 \rightarrow 0.916$ at matched perplexity), suggesting that the MARKTUNE signal is encoded in directions that are partially preserved under output-distribution matching, rather than in surface artifacts that distillation would erase. We caution that this experiment isolates the same-architecture setting and does not test smaller-student distillation, where capacity mismatches between teacher and student may either dampen or amplify the signal. An adversary, lacking the secret key, cannot directly verify whether distillation has removed the watermark; fully training-resistant watermarking, including the smaller-student case, remains an open problem that we discuss further in Appendix G.

Table 23: Same-architecture logit-based distillation: a Llama2-7B student distilled from a MARKTUNE-trained Llama2-7B teacher (**C4-RealNewsLike**). The student inherits substantial detectability without quality loss; the smaller-student case is left to future work.

Method	**AUC↑**	**TPR↑**	**PPL↓**	**Seq-rep-3↓**	**MAUVE↑**
No Watermark	0.500	0.010	$6.55_{1.80}$	0.01	0.98
MARKTUNE (teacher)	0.997	0.982	$6.75_{1.96}$	0.01	0.96
Distilled student	**0.991**	**0.916**	$6.81_{1.92}$	**0.01**	**0.96**

F.8 Watermark reward during training

The rollout-averaged watermark reward during MARKTUNE training is shown below for both model families.



(a) Qwen3-4B-Base (b) Llama2-7B

Figure 7: Watermark reward (GAUSSMARK test statistic) averaged over rollouts during MARKTUNE training.

G Limitations and Future Directions

Theoretical scope. The reverse-KL result in Appendix B.4 analyzes an idealized converged population policy under a linear-softmax model and an explicit strong-concavity assumption. It is not a convergence or detection-power guarantee for the particular finite-step CE-regularized GRPO model. The local analysis in Appendix C complements this result by explaining the quality–reward geometry of CE regularization.

MARKTUNE opens up several promising directions for future work. The first concerns adversaries who go beyond text editing and continued fine-tuning. Although the distillation experiment in Appendix F.7 already shows that the MARKTUNE signal partially survives output-distribution matching, no model-embedded watermark is currently known to remain robust against an adversary who can retrain a model from scratch on the watermarked outputs [24, 27]. A natural next step is to couple MARKTUNE with sharpness-aware optimization [21], steering the watermarked weights toward broad, flat basins of the loss landscape in which parameter-level perturbations (continued fine-tuning, distillation, and partial retraining alike) leave the signal essentially undisturbed. The second direction concerns the choice of detection statistic. The present work instantiates the framework with the GAUSSMARK test statistic and supplies a KGW-reward variant in Appendix E. The same on-policy reward template should extend to semantic-level watermarks such as PMark [36], where the open challenge is to recalibrate the detection-time resampling procedure so that its statistical guarantees continue to hold under the fine-tuned generator. Sample-level correction techniques developed for generative-model sampling, such as soft Metropolis-Hastings reweighting [16], offer a complementary refinement to the on-policy reward estimate. Together, these directions push MARKTUNE toward a watermarking framework that is robust under stronger attack models while applying to a broader class of detection statistics.

H Examples of Generated Texts

In this section, we present sample text completions generated by the Qwen3-4B-Base and Llama2-7B models. Prompts are drawn from the **realnewslike** split of the **C4** dataset and truncated to 200 tokens. For Qwen3-4B-Base, we use a sampling temperature of 0.7 to generate 200-token outputs, and for Llama2-7B, we use a temperature of 1.0. Note that the resulting prompts differ slightly between the two models due to differences in their tokenizers. In the examples below, the text under the heading "Base Model" corresponds to a random completion produced by the original, unwatermarked model, while the text under "Watermarked Model" shows a random completion generated by the model trained with MARKTUNE using the hyperparameters detailed in Appendix D.

H.1 Example 1

Input Prompt: Dubai: The Dubai Financial Market (DFM) witnessed another round of selling on Monday as traders remained cautious ahead of a long holidays. The DFM general index closed 0.44 per cent lower at 2,727.41, as it managed to stay above the support level of 2.706 that has proved critical after the gauged bounced back from these levels twice. Emaar Properties closed 1.2 per cent lower at Dh4.74. Emaar shares are struggling to find a bottom as the stock has been trading near to its previous low. A close over Dh4.85 shall be considered bullish, and ignite recovery to Dh4.95/5.20 in the short term, Shiv Prakash, senior analyst with First Abu Dhabi Bank Securities said in a note. Emaar's movements appear to be concerning at this time of year where investors are expected to be taking positions instead

Qwen3-4B-Base Generations

Base Model: of holding shares ahead of the upcoming holidays, he added. Shares of Emaar had touched Dh4.45 earlier, before recovering to Dh4.50. DFM-listed companies in the real-estate and construction sector declined on the day, with Emaar, Dubai Land Department, Dubai World, Dubai Holding and Dubai Islamic Bank all trading lower. The DFM's general index has been trading below 2.706 since August 12. The stock index had breached this support level in July, as it rallied from its 13-month low of 2.616, which was recorded on July 12. The index has bounced twice after testing the support level. On the other hand, shares of Al Maktoum International Airport had surged 3.8 per cent to Dh1.084 after the airport operator said on Monday that it has entered into a Memorandum of Understanding (MoU) with the International Air

z-score: -0.6454; p-value: 0.7406; perplexity: 5.0251

Watermarked Model: of selling off. We believe the stock has priced in the worst of the recent troubles and given the firm's strong fundamentals, we now expect a bounce towards Dh5.20, he said. We continue to expect Emaar's shares to rally towards Dh5.20 in the next couple of months as the firm has proved resilient in the past. The stock looks cheap given the discount to the firm's intrinsic value, and we now expect the shares to rally towards Dh5.20 in the next couple of months, he added. The real estate sector was the biggest loser of the day, with the sector index down by 1.2 per cent at 3,124.13. Almarai (ALM) declined 2.2 per cent to Dh5.85, while Alrosa (ARO) shed 1.9 per cent to Dh1.53 and Alshaya (ALY) dropped 1.

z-score: 4.5114; p-value: 3.2187e-06; perplexity: 4.625

Input Prompt: Dubai: The Dubai Financial Market (DFM) witnessed another round of selling on Monday as traders remained cautious ahead of a long holidays. The DFM general index closed 0.44 per cent lower at 2,727.41, as it managed to stay above the support level of 2.706 that has proved critical after the gauged bounced back from these levels twice. Emaar Properties closed 1.2 per cent lower at Dh4.74. Emaar shares are struggling to find a bottom as the stock has been trading near to its previous low. A close over Dh4.85 shall be considered bullish, and

ignite recovery to Dh4.95/5.20 in the short term, Shiv Prakash, senior analyst with First Abu Dhabi Bank Securities

Llama2-7B Generations

Base Model: (FASB Securities), said.The stock has traded in a narrow range for almost two years, during which time the stock did gain some momentum but failed to penetrate the key resistance point of Dh4.98/5. Since then, there has been no meaningful progress and the stock price has been in a free fall mode, he said.Emirates Telecommunication Group was the most active stock traded on the index with nearly 9.619 million shares, closely followed by Arabtec with over 8.219 million shares.On value terms, Emaar Properties contributed roughly Dh153.32 million to the market capitalisation, followed by Arabtec with Dh43.25 million and Emirates NBD with Dh20.69 million.In sectoral terms, the real estate sector led the top gainers with a surge

z-score: 0.5016; p-value: 0.3079; perplexity: 5.2253

Watermarked Model: (FASB Securities), told the Gulf News.Dubai Islamic Bank, meanwhile, was another leading index heavyweight to gain ground on Monday. Shares of the bank picked up 0.77 per cent to close at Dh6.42.On the DFM and other markets in the region, we're looking at the 19th [August] as a down day for traders and investors, said Ahmed Saqr, assistant director, financial control department at Abu Dhabi Islamic Bank.After the Eid week, many traders and investors will have just a few days left in August to trade, he added.Investors will take a wait-and-watch approach, he said.Their sentiment towards the markets may turn negative at the beginning of September, because of the US September rates hike talks. This could negatively

z-score: 4.3386; p-value: 7.1526e-06; perplexity: 5.4523

H.2 Example 2

Input Prompt: MADRID: As Cristiano Ronaldo enjoyed his time off at the French Riviera, his Portugal team-mates were busy handing Italy another disappointing result in the UEFA Nations League. Portugal did not need Ronaldo, who skipped international duty to recharge after a busy summer, to beat Italy 1-0 in Europe's newest soccer competition. Elsewhere, Turkey mounted a remarkable comeback from two goals down against Sweden, while Kosovo made history with their first competitive win since being admitted to UEFA in 2016.The European champions easily outplayed Italy, which had opened its Nations League campaign with a disappointing 1-1 home draw against Poland in its first competitive match under coach Roberto Mancini. Andre Silva scored on a counterattack early in the second half to give Portugal the Group 3 win in the top-tier League A. The hosts had several chances to extend their lead at the Stadium of Light in Lisbon, while Italy created little and relied mostly on set pieces. The Nations League gives

Qwen3-4B-Base Generations

Base Model: a fresh impetus to the tournament that replaced the European Championship qualifiers. The top-tier league is divided into four pots, with each group consisting of four teams. The top two teams in each group will qualify for the 2020 European Championship in France. Turkey, which is not currently ranked in the top 20, came back from two goals down to beat Sweden 3-2 at the Group 3 venue in Istanbul. The Swedes took the lead in the 25th minute through a penalty, but the hosts had already equalised a minute later on a goal from a corner. Turkey took the lead through a goal from Ali Kocak, but the Swedes equalised through a Hakan Yakin penalty, before Kocak scored in the 70th minute to make the match up until then a 2-2 stalemate. The hosts took the lead through a goal from Cemre Onay in the 89th minute to win

z-score: 0.9452; p-value: 0.1722; perplexity: 4.9626

Watermarked Model: the teams a chance to warm up for next summer's Euros, where they will face each other. But with Italy's defence vulnerable, the likes of Paulo Dybala and the injured Mario Balotelli were hardly needed. We should have won the game, but we didn't get the best of it, Mancini said after the defeat. This was a good way to start the Nations League, but we now have to find a way to get better because it's not enough to beat Portugal. Italy's next opponents in the group stage are the likes of Poland and Wales, but the 2020 European champions are already planning for the big games after the Nations League ends. We'll start preparing for the next phase of the year, because we know we're going to face some tough opponents in the next few weeks, Mancini said. We'll focus on those games and try to make the most of the opportunities we have. The Italians will also be looking to bounce back

z-score: 4.3033; p-value: 8.4043e-06; perplexity: 5.4245

Input Prompt: MADRID: As Cristiano Ronaldo enjoyed his time off at the French Riviera, his Portugal team-mates were busy handing Italy another disappointing result in the UEFA Nations League. Portugal did not need Ronaldo, who skipped international duty to recharge after a busy summer, to beat Italy 1-0 in Europe's newest soccer competition. Elsewhere, Turkey mounted a remarkable comeback from two goals down against Sweden, while Kosovo made history with their first competitive win since being admitted to UEFA in 2016. The European champions easily outplayed Italy, which had opened its Nations League campaign with a disappointing 1-1 home draw against Poland in its first competitive match under coach Roberto Mancini. Andre Silva scored on a counterattack early in the second half to give Portugal the Group 3 win in the top-tier League A. The hosts

Llama2-7B Generations

Base Model: dominated for most of the game by playing Portugal more openly and allowing more space in midfield, but they struggled to take advantage. Portugal's previous victory over Italy was 1-0 in the Euro 2016 semifinals when it lifted the tournament's trophy. Italy played most of that game with 10 men after defender Giorgio Chiellini received a red card. Portugal was the only one of the 21 European members of the competition which had already played two games in Group 3 prior to the opener against Italy. The host nation, Spain and Poland had one game apiece and will continue the campaign this week. Italy lost its first two fixtures in the Nations League and missed out on qualification for the 2018 World Cup in Russia, while Portugal is one of Europe's major powers. It finished third in its group at the World Cup, losing

z-score: -1.7278; p-value: 0.9579; perplexity: 5.8205

Watermarked Model: dominated for most of the game by patiently counterattacking when Italy tried to play with the ball. The Portuguese had their most threatening chances when Matteo Politano was sent off after 30 minutes for a rash challenge on Joao Mario outside the penalty area. Ricardo Horta shot wide from a good position. Italy dominated ball play in the midfield and defense after Portugal scored in the 53rd, but lacked the final attacking touch. Portugal, which lost its top spot in the FIFA rankings last week, remained two points above Sweden in Group 2 after both teams began their campaigns. The Portuguese host Sweden in September, while Italy travels to Poland. Sweden opened the match with two goals from former Premier League player Emil Forsberg, who scored in the first and 55th minutes. Sweden looked to have the title in its sights as they pressed for a

z-score: 3.5713; p-value: 0.0002; perplexity: 6.2521

H.3 Example 3

Input Prompt: The New York Times on Thursday printed a special section featuring portraits of nearly every member of the record number of female lawmakers in Congress. A viral Twitter thread from the Times' design editor, Josh Crutchmer, includes a video of the covers the night before publication that has been viewed over 320,000 times as of Thursday morning. The photos, which were taken on Capitol Hill over five days by Times photographers Elizabeth D. Herman and Celeste Sloman, are meant as a testament to what power looks like in 2019, according to the paper. Like the work of Kehinde Wiley, who painted Barack Obama Barack Hussein ObamaChina, Russia, Iran rise in Latin America as US retreats Castro wants to follow Obama's lead on balancing presidency with fatherhood Trump's regulatory rollback boosts odds of a financial crisis MORE's official presidential portrait, these photographs evoke the imagery we are used to seeing in the halls of power, but place people not previously seen

Qwen3-4B-Base Generations

Base Model: occupying them.T̈he photographs are intended to show the power of women in our government, according to a Times source. The Times' print section, which ran the portraits on Thursday, also printed an article titled The 2020 Women's History Month Celebration. The article contains a list of female lawmakers and the causes they represent. The Times also published a companion piece featuring portraits of 12 women in Congress who have taken on more than just their usual roles. The article also included a video of the portraits, which had been viewed over 350,000 times as of Thursday morning, and a list of the women whose portraits appeared in the print section. The women profiled in the article include Rep. Jackie Speier, a California Democrat who has been battling cancer since 2015 and who, as of 2019, is one of only 13 women in Congress who have survived cancer. The article also features Rep. Barbara

z-score: 0.6038; p-value: 0.2729; perplexity: 5.9870

Watermarked Model: in that position. It's the kind of thing the paper has been doing for decades. In 2016, it published a series of its own power portraits featuring the leaders of every major political party during the presidential primaries. This year, the paper has been highlighting the stories of the women running for president and the women in the other major races. But this year's section is particularly special because it's the first time the paper has managed to get the coverage it deserves for its work. The section is the culmination of a massive effort by the paper's editorial and design teams to make the most of the historic moment. There are so many great stories in this election cycle, but we wanted to make sure we told the ones that matter most, said the paper's executive editor, Dean Baquet. The section is just the beginning of what the paper hopes will be a long-term commitment to telling the stories of women in politics. The paper has already announced plans to launch a new

z-score: 3.5408; p-value: 0.0002; perplexity: 5.0692

Input Prompt: The New York Times on Thursday printed a special section featuring portraits of nearly every member of the record number of female lawmakers in Congress. A viral Twitter thread from the Times' design editor, Josh Crutchmer, includes a video of the covers the night before publication that has been viewed over 320,000 times as of Thursday morning. The photos, which were taken on Capitol Hill over five days by Times photographers Elizabeth D. Herman and Celeste Sloman, are meant as a testament to what power looks like in 2019, according to the paper. Like the work of Kehinde Wiley, who painted Barack Obama Barack Hussein ObamaChina, Russia, Iran rise in Latin America as US retreats Castro wants to follow Obama's lead on balancing presidency with fatherhood Trump's regulatory rollback boosts odds

Llama2-7B Generations

Base Model: of successful Obamacare repeal MORE's official presidential portrait, our intent is to show just how diverse the new Congress is from all perspectives, Crutchmer tweeted. According to The Washington Post, the photos also include Rep. Alexandria Ocasio-Cortez, as well as the first Muslim woman elected to Congress. Rep. Ayanna Pressley, who won her election in Massachusetts and is the seat Ocasio-Cortez will take, is featured as the main photo on the cover, which reads Unity in Action. Rep. Ilhan Omar on Monday became the first Somali-American and one of the first Muslim women to serve in Congress. Lawmakers said upon arriving in Washington that they would be a force multiplier, as Omar said. In a statement to The Hill when photos first came out, Omar said the group is going to be

z-score: 0.3792; p-value: 0.3522; perplexity: 6.2308

Watermarked Model: of nuclear incident MORE for the first time, this project honors women on the national stage, portraying them as they are, Times executive editor Dean Baquet said in a statement. One of the lawmakers, Rep. Deb Haaland Deb HaalandNational Archives to release Trump's passport application, other records on Friday: report Judy Chu is the next chairwoman of the Congressional Asian Pacific American Caucus Nadler's top priority is fighting 'epidemic of corruption in our political system' MORE (D-N.M.) who is the first Native American woman elected to Congress, said the photoshoot reflected the way women have been portrayed for centuries, both in media and in the public realmB̈aquet was in charge of the paper's decision to use the photo of Rep. Alexandria Ocasio-Cortez (D-N.Y.) in a sk

z-score: 3.6125; p-value: 0.0001; perplexity: 6.2042

Table 24: Extension of Table 2 with perplexity (PPL) to validate attack plausibility.

Attack	Metric	KGW-1	KGW-2	Gumbel-max	SynthID	KGW-D	GaussMark	MarkTune
Qwen3-4B-Base, T=0.7								
Del-0.2	**AUC↑**	0.881	0.975	0.993	0.987	0.947	0.908	0.921
	TPR↑	0.330	0.746	0.948	0.890	0.582	0.366	0.484
	PPL↓	20.50	25.19	18.82	19.15	21.23	19.49	19.39
Del-0.5	**AUC↑**	0.617	0.697	0.778	0.780	0.851	0.622	0.744
	TPR↑	0.040	0.082	0.162	0.094	0.196	0.048	0.068
	PPL↓	98.53	117.25	94.82	94.19	99.95	96.82	94.79
Sub-0.2	**AUC↑**	0.873	0.986	0.993	0.986	0.939	0.919	0.948
	TPR↑	0.324	0.826	0.922	0.846	0.578	0.384	0.552
	PPL↓	24.71	28.64	21.46	23.06	25.44	23.42	23.17
Sub-0.5	**AUC↑**	0.638	0.747	0.840	0.834	0.837	0.775	0.783
	TPR↑	0.048	0.106	0.348	0.182	0.190	0.100	0.110
	PPL↓	71.05	80.28	60.34	67.19	72.48	68.96	67.52
Dipper-1	**AUC↑**	0.771	0.884	0.945	0.941	0.855	0.915	0.977
	TPR↑	0.142	0.372	0.748	0.576	0.274	0.438	0.852
	PPL↓	6.51	7.30	6.07	6.18	6.56	6.24	6.16
Dipper-2	**AUC↑**	0.573	0.616	0.721	0.700	0.589	0.795	0.851
	TPR↑	0.026	0.038	0.146	0.074	0.030	0.138	0.254
	PPL↓	7.72	8.34	7.81	7.59	7.70	7.80	7.62
Translate	**AUC↑**	0.831	0.928	0.955	0.968	0.912	0.948	0.973
	TPR↑	0.242	0.552	0.846	0.818	0.404	0.574	0.858
	PPL↓	9.16	11.37	8.33	8.73	9.29	8.95	8.81
Llama2-7B, T=1.0								
Del-0.2	**AUC↑**	0.853	0.975	0.993	0.979	0.914	0.835	0.928
	TPR↑	0.232	0.712	0.924	0.774	0.410	0.242	0.506
	PPL↓	23.71	27.13	21.66	22.55	23.54	22.86	22.49
Del-0.5	**AUC↑**	0.581	0.711	0.790	0.777	0.803	0.617	0.783
	TPR↑	0.024	0.060	0.224	0.114	0.148	0.036	0.080
	PPL↓	92.81	105.16	87.05	88.56	95.12	89.98	88.95
Sub-0.2	**AUC↑**	0.878	0.980	0.993	0.987	0.887	0.853	0.944
	TPR↑	0.300	0.756	0.950	0.862	0.406	0.228	0.578
	PPL↓	22.97	25.89	21.42	22.50	22.84	22.68	22.53
Sub-0.5	**AUC↑**	0.638	0.772	0.862	0.843	0.759	0.759	0.809
	TPR↑	0.030	0.126	0.296	0.140	0.144	0.098	0.122
	PPL↓	48.50	52.78	44.80	46.69	47.12	46.78	46.51
Dipper-1	**AUC↑**	0.757	0.862	0.950	0.943	0.814	0.857	0.961
	TPR↑	0.116	0.322	0.700	0.552	0.238	0.240	0.836
	PPL↓	8.33	8.91	7.90	8.11	8.19	8.13	8.09
Dipper-2	**AUC↑**	0.551	0.614	0.711	0.710	0.590	0.687	0.859
	TPR↑	0.014	0.032	0.148	0.084	0.026	0.060	0.270
	PPL↓	9.34	9.58	8.64	8.97	9.25	9.07	9.02
Translate	**AUC↑**	0.777	0.922	0.961	0.949	0.889	0.925	0.966
	TPR↑	0.146	0.448	0.796	0.612	0.324	0.436	0.834
	PPL↓	13.27	14.99	11.69	12.68	12.86	12.56	12.47

I AI Use Statement

We used LLM-based tools for manuscript editing and literature-review assistance, and for implementation, refactoring, and documentation of the released code. All AI-assisted manuscript edits were reviewed by the authors for accuracy, appropriate attribution, and consistency with the underlying research. We take full responsibility for the final manuscript and released implementation, including the mathematical claims, proofs, experimental results, and analysis.

J Ethics Statement

The research presented in this work adheres to the NeurIPS Code of Ethics. This paper does not involve crowdsourcing or research with human subjects, and introduces no new datasets containing personal data. MARKTUNE aims to support provenance attribution, accountability for misuse, and intellectual-property protection for open-weight language models, where users can modify the decoding pipeline. Watermark detection nevertheless carries risks of false attribution and missed detection, and robustness is limited to the conditions evaluated in this work. Detection outcomes should therefore be interpreted with their statistical uncertainty and should not serve as the sole basis for punitive decisions or claims of ownership.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The method and theoretical claims are developed in Section 3 and Appendices B and C. Section 4 supports the reported quality–detectability trade-off, downstream utility, cross-dataset generalization, and robustness under the evaluated settings.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix G discusses the scope of the population and local analyses, stronger attack models, and extensions to other detection statistics. Section 4 and Appendix F specify the evaluated models, datasets, and attack settings.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Propositions 3.1 and 3.3, Proposition B.1, and Theorems B.2 and B.3 are stated with their assumptions and proved in Appendix B. The reverse-KL power result concerns an idealized population policy, while Appendix C gives the assumptions and derivations for the local CE analysis.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 and Appendices A and D specify the training and evaluation protocols. The public implementation supplies model-specific configurations and data-preparation, training, and evaluation commands for the main experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code, model-specific configurations, and installation and execution instructions are available at https://github.com/zhyzmath/MarkTune-LLM-Watermarking.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 and Appendix D describe the training and evaluation settings, including the per-model hyperparameters in Tables 3 and 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1 reports perplexity means and standard deviations across 500 evaluation prompts. Figures 3 and 6 report bootstrap-estimated 90% confidence intervals (500 resamples) on TPR@1%FPR, with the procedure described in Appendix D. These summaries quantify variation across evaluation samples.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer Yes if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix D.6 reports training on 4 NVIDIA H200 GPUs for approximately 8 h (Qwen3-4B-Base) and 13.5 h (Llama2-7B). Appendix D.4 specifies the single-H200 latency benchmark, and Appendix D.5 reports detector access and peak GPU memory requirements.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Appendix J discusses conformity with the NeurIPS Code of Ethics and the responsible interpretation of detection results. No human subjects or new personal-data datasets are involved.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix J discusses provenance attribution, misuse accountability, and intellectual-property protection, alongside false attribution, missed detection, and the risks of using watermark evidence as the sole basis for punitive decisions or ownership claims.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The public repository distributes code and configurations, with no model weights or datasets. Pretrained models and datasets are obtained separately from their original providers under the applicable access conditions; Appendix J discusses responsible use of detection results.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets, pretrained models, and baseline methods are credited in Section 4 and Appendix D. The repository's `THIRD_PARTY.md` identifies upstream software sources and versions, and the bundled license files retain the Apache-2.0 terms for these components. Model weights and datasets are obtained from their original providers under their respective licenses and access conditions.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The public repository includes a README with installation, data preparation, and training/evaluation instructions, together with model-specific configurations. The `LICENSE` and `THIRD_PARTY.md` files document the MIT license for MarkTune and the source attributions and licenses of third-party components.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM-based assistants were used for writing, literature review, and preparation of the released code, as disclosed in Appendix I. These assistants are not components of the watermarking or evaluation algorithms.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.