THEORETICALLY GROUNDED FRAMEWORK FOR LLM WATERMARKING: A DISTRIBUTION-ADAPTIVE APPROACH

Haiyun He[1]* **Yepeng Liu** [2]* **Ziqiao Wang**[3]† **Yongyi Mao**[4] **Yuheng Bu**[2]‡

[1]Cornell University [2]University of Florida [3]Tongji University [4]University of Ottawa

hh743@cornell.edu, {yepeng.liu, buyuheng}@ufl.edu, ziqiaowang@tongji.edu.cn, ymao@uottawa.ca

ABSTRACT

Watermarking has emerged as a crucial method to distinguish AI-generated text from human-created text. In this paper, we present a novel theoretical framework for watermarking Large Language Models (LLMs) that jointly optimizes both the watermarking scheme and the detection process. Our approach focuses on maximizing detection performance while maintaining control over the worst-case Type-I error and text distortion. We characterize *the universally minimum Type-II error*, showing a fundamental trade-off between watermark detectability and text distortion. Importantly, we identify that the optimal watermarking schemes are adaptive to the LLM generative distribution. Building on our theoretical insights, we propose an efficient, model-agnostic, distribution-adaptive watermarking algorithm, utilizing a surrogate model alongside the Gumbel-max trick. Experiments conducted on Llama2-13B and Mistral-8$\times$7B models confirm the effectiveness of our approach. Additionally, we examine incorporating robustness into our framework, paving the way for future watermarking systems that withstand adversarial attacks more effectively.

1 INTRODUCTION

Arising with Large Language Models (LLMs) (Touvron et al., 2023) that demonstrate stunning power are substantial risks: spreading disinformation, generating fake news, engaging in plagiarism, etc. Such risks elevate as LLMs are increasingly widely adopted for content generation. Distinguishing AI-generated content from human-written text is then critically demanded and watermarking serves as an effective solution to address this challenge.

Existing watermarking techniques for AI-generated text can be classified into two categories: post-process and in-process. Post-process watermarks (Brassil et al., 1995; Yoo et al., 2023; Yang et al., 2023; Munyer et al., 2023; Yang et al., 2022; Sato et al., 2023; Zhang et al., 2024; Abdelnabi & Fritz, 2021) are applied after the text is generated, while in-process watermarks (Wang et al., 2023; Fairoze et al., 2023; Hu et al., 2023; Huo et al., 2024; Zhang et al., 2023; Tu et al., 2023; Ren et al., 2023) are embedded during generation. Between the two types, in-process watermarking is more favorable due to its flexibility and numerous techniques have been proposed to seamlessly integrate watermarks into the generation process. Notably, an ideal in-process watermarking scheme for LLMs should have four desired properties: 1) **Detectability**: the watermarking can be reliably detected with Type-I error controlled; 2) **Distortion-free** (Christ et al., 2024; Kuditipudi et al., 2023): the watermarked text preserves the quality of the original generated text by maintaining the original text distribution; 3) **Robustness** (Zhao et al., 2023; Liu & Bu, 2024): the watermark is resistant to modifications aimed at its removal; 4) **Model-agnostic** (Kirchenbauer et al., 2023a): detection does not require knowledge of the original watermarked LLMs or the prompts. Clearly, one expects tension between these dimensions. Yet, despite the great efforts in designing watermarking and detection schemes that heuristically balance these factors, theoretical understanding of the fundamental trade-offs therein is rather limited to date.

Among existing theoretical analyses, Huang et al. (2023) frame statistical watermarking as a test of independence between the text and the watermark, and analyze the optimal watermarking scheme

*Equal contribution.

†This work was carried out while the author was a PhD student at University of Ottawa.

‡Corresponding author.

for a specific detection process. While their analysis can be extended to a model-agnostic setting, they do not propose a practical algorithm. In contrast, Li et al. (2024) propose a surrogate hypothesis testing framework based on i.i.d. pivotal statistics, with the goal of identifying statistically optimal detection rules for a given watermarking scheme. However, their method depends on a suitable but unoptimized watermarking scheme, and their detection rule is not necessarily optimal for the original independence test. While these studies provide useful insights, they fall short of capturing the jointly optimal watermarking scheme and detection rule, limiting their practicality and effectiveness in real-world applications.

In this paper, we formulate the LLM watermarking problem as an independence test between the text and an auxiliary variable to jointly optimize *both* the watermarking scheme and the detector. Unlike the classical watermarking paradigm, which employs fixed watermark messages or distributions, this study investigates a broader definition of watermarking where watermarks are adaptively generated based on the generative distribution of LLMs. This approach enables the full exploitation of the generative capability of LLMs, thereby enhancing watermark detection performance to new limits (Figure 1). Our theoretical framework characterizes the fundamental trade-off between detectability, distortion, and robustness by minimizing Type-II error. To capture the **detectability**, we define universal optimality in two aspects: 1) controlling the false alarm rate across *all* possible text distributions, and 2) obtaining a universally minimum Type-II error for *all* possible detectors and watermarking schemes. Additionally, we measure the **distortion** of a watermarked LLM using the divergence between the watermarked text distribution and the original text distribution. **Robustness**, on the other hand, depends on modifications to the watermarked text, such as replacement, deletion/insertion, or paraphrasing. Unlike existing approaches that evaluate robustness via experiments (Liu & Bu, 2024) or provide detection error bounds under specific modifications (Kuditipudi et al., 2023; Zhao et al., 2023), our framework covers a broader range of potential attacks, including those that preserve the semantics of the text.

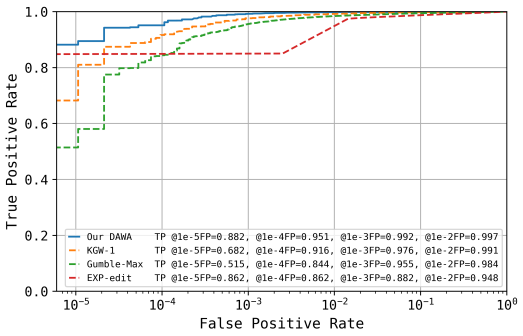

Figure 1: Comparison of TPR at extremely low FPR among different watermarking methods.

Our contributions can be summarized as follows:

- In Section 2, we propose a theoretical framework for LLM watermarking and detection that encompasses most modern LLM watermarking methods. This framework features a common randomness shared between watermark generation and detection to perform an independence test.
- In Section 3, we characterize the universally minimum Type-II error in our framework, revealing a fundamental trade-off between detectability and distortion. More importantly, we identify the class of jointly optimal detectors and watermarking schemes, providing a guideline for practical design, namely, optimal watermarking schemes should adapt to the generative distribution of LLMs.
- In Section 4, we introduce a practical token-level optimal watermarking scheme that guarantees detection performance and demonstrates inherent robustness against token replacement attacks. In Section 5, we present a novel watermarking method, DAWA (**D**istribution-**A**daptive **W**atermarking **A**lgorithm), which leverages a surrogate language model and the Gumbel-max trick to achieve *model-agnosticism* and *computational efficiency*.
- We perform extensive experiments (Section 6) using various language models, including Llama2-13B (Touvron et al., 2023) and Mistral-8$\times$7B (Jiang et al., 2023), across multiple datasets. DAWA is shown to consistently outperform the compared methods, demonstrating robust performance against token replacement attacks. As shown in Figure 1, DAWA achieves superior detection capabilities at extremely low false positive rates.
- Lastly, we explore how to incorporate *robustness* against semantic-invariant attacks into our theoretical framework (Appendix L), providing insights for designing optimal *semantic-based* watermarking systems that are robust to such attacks.

Other Related Literature. The advance of LLMs boosts productivity but also presents challenges like bias and misuse. Watermarking addresses these challenges by tracing AI-generated content and distinguishing it from human-created material. Many watermarking methods for LLMs have been

proposed (Zhou et al., 2024; Fu et al., 2024; Giboulot & Furon, 2024; Wu et al., 2023; Kirchenbauer et al., 2023b), including biased and unbiased (distortion-free) watermarking. Biased watermarks (Kirchenbauer et al., 2023a; Zhao et al., 2023; Liu & Bu, 2024; Liu et al., 2024; Qu et al., 2024) typically alter the next-token prediction distribution to increase the likelihood of sampling certain tokens. For example, Kirchenbauer et al. (2023a) divides the vocabulary into green and red lists and slightly enhances the probability of green tokens in the next token prediction (NTP) distribution. Unbiased watermarks (Zhao et al., 2024; Fernandez et al., 2023; Boroujeny et al., 2024; Christ et al., 2024; Giboulot & Furon, 2024) maintain the original NTP distributions or texts unchanged, using various sampling strategies to embed watermarks. The Gumbel-max watermark (Aaronson, 2023) utilizes the Gumbel-max trick (Gumbel, 1954) to sample the next token, while Kuditipudi et al. (2023) introduces an inverse transform method for this purpose.

Most existing watermarking schemes and detectors are heuristic and lack theoretical support. Traditional post-process watermarking schemes, which apply watermarks after generation, have been extensively studied from information-theoretic perspective (Martinian et al., 2005; Moulin & O'Sullivan, 2000; Chen, 2000; Merhav & Ordentlich, 2006; Merhav & Sabbag, 2008). For in-process watermarking, while two prior works (Huang et al., 2023; Li et al., 2024) attempt to derive theoretically optimal schemes or detectors, their solutions are either not jointly optimized or lack universal optimality as achieved in our paper. In contrast, we propose a framework that jointly optimizes the watermarking scheme and detector for an optimal configuration of both components.

2 Preliminaries and Problem Formulation

Notations. For any set $\mathcal{X}$, we denote the space of all probability measures over $\mathcal{X}$ by $\mathcal{P}(\mathcal{X})$. For a sequence of random variables $X_1, X_2, \ldots, X_n$, and any $i, j \in [n]$ with $i \leq j$, we denote $X_i^j := (X_i, \ldots, X_j)$. We may use distortion function, namely, a function $\mathsf{D} : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to [0, +\infty)$ to measure the dissimilarity between two distributions in $\mathcal{P}(\mathcal{X})$. For example, the total variation distance, as a distortion, between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is $\mathsf{D}_{\mathsf{TV}}(\mu, \nu) := \int \frac{1}{2} |\frac{d\mu}{d\nu} - 1| \, d\nu$. For any set $A \subseteq \mathcal{X}$, we use δ_A to denote its indicator function, namely, $\delta_A(x) := \mathbb{1}\{x \in A\}$. Additionally, we denote $(x)_+ := \max\{x, 0\}$ and $x \wedge y := \min\{x, y\}$.

Tokenization and NTP. LLMs process text through "tokenization," namely, breaking it down into words or word fragments called "tokens." An LLM generates text token by token. Let $\mathcal{V}$ denote the token vocabulary, typically of size $|\mathcal{V}| = \mathcal{O}(10^4)$ (Liu, 2019; Radford et al., 2019; Zhang et al., 2022; Touvron et al., 2023). An *unwatermarked* LLM generates the next token X_t based on a prompt pt and the previous tokens x_1^{t-1} by sampling the Next-Token Prediction (NTP) distribution $Q_{X_t|x_1^{t-1},\mathrm{pt}}$. For simplicity, the prompt dependency is suppressed in notation throughout the paper. The joint distribution of a length-T generated token sequence X_1^T is then given by $Q_{X_1^T} := \prod_{t=1}^T Q_{X_t|X_1^{t-1}}$, which we assume to be identical to one that governs the human-generated text.

A Framework for Watermarking Scheme. Traditional post-hoc detectors identify AI-generated text by dividing the entire text space into rejection and acceptance regions, which relies on the assumption that certain sentences cannot be produced by humans. In contrast, modern LLM watermarking schemes achieve the same goal by analyzing the dependence structure between text X_1^T and an auxiliary random sequence ζ_1^T, thereby avoiding this unrealistic assumption.

In this paper, we propose a general framework for LLM watermarking and detection, as shown in Figure 2, which encompasses most of the existing watermarking schemes. The watermarking scheme and detector share a common randomness represented by an auxiliary random sequence ζ_1^T drawn from a space $\mathcal{Z}^T$ (either discrete or continuous). After passing through a watermarking scheme, the watermarked LLM samples token sequence according to the modified NTP distribution $P_{X_t|x_1^{t-1},\zeta_1^T}$, where $P_{X_1^T|\zeta_1^T} = \prod_{t=1}^T P_{X_t|X_1^{t-1},\zeta_1^T}$. This process associates the generated text X_1^T with an auxiliary sequence ζ_1^T. Thus, the joint distribution of the watermarked token sequence X_1^T is $P_{X_1^T}$, which might be different

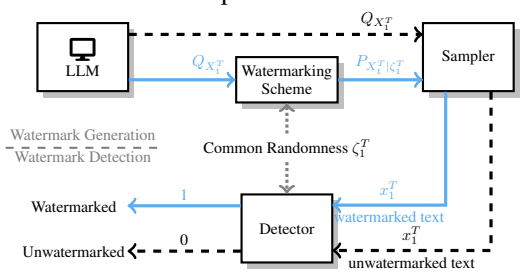

Figure 2: Overview of LLM watermarking and detection.

from the original $Q_{X_1^T}$. The detector can then distinguish whether the received sequence X_1^T is watermarked or not based on the common randomness.

To evaluate the *distortion level* of a watermarking scheme, we measure the statistical divergence between the watermarked text distribution $P_{X_1^T}$ and the original one $Q_{X_1^T}$.

Definition 1 (ϵ-distorted watermarking scheme)**.** *A watermarking scheme is* ϵ-distorted *with respect to distortion* D*, if* $\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$*. Here,* D *can be any distortion metric.*

Common examples of such divergences include squared distance, total variation, KL divergence, and Wasserstein distance. For $\epsilon = 0$, the watermarking scheme is unbiased or distortion-free.

Specifically, our formulation allows the auxiliary random sequence ζ_1^T to take an arbitrary structure, which contrasts the rather restricted i.i.d. assumption considered in Li et al. (2024, Working Hypothesis 2.1). In practice, ζ_1^T is usually randomly generated using a shared `key` accessible during both watermark generation and detection. At first glance, our formulation may appear abstract, but its flexibility enables existing watermarking schemes to be interpreted as special cases within this framework.

Example 1 (Existing watermarking schemes as special cases)**.** *In the Green-Red List watermarking scheme (Kirchenbauer et al., 2023a), at each position* t*, the vocabulary* $\mathcal{V}$ *is randomly split into a green list* $\mathcal{G}$ *and a red list* $\mathcal{R}$*, with* $|\mathcal{G}| = \rho|\mathcal{V}|$ *for some* $\rho \in (0,1)$*. This split is represented by a* $|\mathcal{V}|$*-dimensional binary auxiliary variable* ζ_t*, indexed by* $x \in \mathcal{V}$*, where* $\zeta_t(x) = 1$ *means* $x \in \mathcal{G}$*; otherwise,* $x \in \mathcal{R}$*. The watermarking scheme is as follows:*

- *Compute a hash of the previous token* X_{t-1} *using a hash function* $h : \mathcal{V} \times \mathbb{R} \to \mathbb{R}$ *and a shared secret* `key`*, i.e.,* $h(X_{t-1}, \texttt{key})$*.*
- *Use* $h(X_{t-1}, \texttt{key})$ *as a seed to uniformly sample the auxiliary variable* ζ_t *from the set* $\{\zeta \in \{0,1\}^{|\mathcal{V}|} : \|\zeta\|_1 = \rho|\mathcal{V}|\}$ *to construct the green list* $\mathcal{G}$*.*
- *Sample* X_t *from the adjusted NTP distribution which increases the logit of tokens in* $\mathcal{G}$ *by* $\delta > 0$*:*

$$P_{X_t|x_1^{t-1},\zeta_t}(x) = \frac{Q_{X_t|x_1^{t-1}}(x)\exp(\delta \cdot \mathbb{1}\{\zeta_t(x)=1\})}{\sum_{x\in\mathcal{V}} Q_{X_t|x_1^{t-1}}(x)\exp(\delta \cdot \mathbb{1}\{\zeta_t(x)=1\})}.$$

A discussion of how our formulation encompasses several other schemes is provided in Appendix B.

Hypothesis Testing for Watermark Detection. Note that a sequence X_1^T generated by a watermarked LLM depends on ζ_1^T, while X_1^T and ζ_1^T are independent if written by humans. Therefore, detection involves distinguishing the following two hypotheses based on the pair (X_1^T, ζ_1^T):

- H_0: X_1^T is generated by a human, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$;
- H_1: X_1^T is generated by a watermarked LLM, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T,\zeta_1^T}$.

We consider a model-agnostic detector $\gamma : \mathcal{V}^T \times \mathcal{Z}^T \to \{0,1\}$, which maps (X_1^T, ζ_1^T) to the hypothesis index (see Figure 2). In theory, we assume that the auxiliary sequence ζ_1^T can be fully recovered from X_1^T and the common randomness, while this assumption is dropped in practice.

Detection performance is measured by the Type-I (false alarm) and Type-II (missed detection) error probabilities:

$$\beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \coloneqq \Pr(\gamma(X_1^T, \zeta_1^T) \neq 0 \mid \mathrm{H}_0), \;\; \beta_1(\gamma, P_{X_1^T,\zeta_1^T}) \coloneqq \Pr(\gamma(X_1^T, \zeta_1^T) \neq 1 \mid \mathrm{H}_1). \quad (1)$$

Optimization Problem. Given that human-generated texts can vary widely, within our proposed framework, we aim to control the *worst-case* Type-I error $\sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T})$ at a given $\alpha \in (0,1)$ while minimizing Type-II error. Our objective is to design an ϵ-distorted watermarking scheme and a model-agnostic detector by solving the following optimization:

$$\inf_{\gamma, P_{X_1^T,\zeta_1^T}} \beta_1(\gamma, P_{X_1^T,\zeta_1^T}) \qquad \text{s.t. } \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \leq \alpha,\ \mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon. \quad \text{(Opt-O)}$$

The optimal objective value, denoted as $\beta_1^*(Q_{X_1^T}, \alpha, \epsilon)$, is termed as *universally minimum Type-II error*. This universality is due to its applicability across all potential detectors and watermarking schemes, as well as its validity under the worst-case Type-I error scenario.

3 JOINTLY OPTIMAL WATERMARKING SCHEME AND DETECTOR

In this section, we aim to solve the optimization in (Opt-O) and identify the jointly optimal watermarking scheme and detector. However, solving (Opt-O) is challenging due to the binary nature of γ and the vast set of possible γ, sized $2^{|\mathcal{V}|^T|\mathcal{Z}|^T}$. To address this, we begin with a fixed $\gamma(X_1^T, \zeta_1^T) = \mathbb{1}\{(X_1^T, \zeta_1^T) \in \mathcal{A}_1\}$, where $\mathcal{A}_1$ defines the acceptance region for H_1, aiming to uncover a potential structure for the optimal detector. To this end, we simplify (Opt-O) as

$$\inf_{P_{X_1^T,\zeta_1^T}} \beta_1(\gamma, P_{X_1^T,\zeta_1^T}) \qquad \text{s.t. } \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha,\ \mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon. \tag{Opt-I}$$

Error-Distortion Tradeoff. We first derive a lower bound for the minimum Type-II error in (Opt-I), which surprisingly does not depend on the selected detector γ and therefore also applies to (Opt-O). We then pinpoint a type of detector and watermarking scheme that attains this lower bound, indicating that it represents the universally minimum Type-II error. Thus, the proposed detector and watermarking scheme are jointly optimal, as detailed in Theorem 2. The theorem below establishes this universally minimum Type-II error for all feasible watermarking schemes and detectors.

Theorem 1 (Universally minimum Type-II error)**.** *The universally minimum Type-II error attained from* (Opt-O) *is*

$$\beta_1^*(Q_{X_1^T}, \alpha, \epsilon) = \min_{P_{X_1^T}:\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon} \sum\nolimits_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+, \tag{2}$$

which is achieved by the watermarked distribution

$$P^*_{X_1^T} = \operatorname*{arg\,min}_{P_{X_1^T}:\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+. \tag{3}$$

By setting D *as total variation distance* D_{TV}*,* (2) *can be simplified as follows: If* $\sum_{x_1^T}(\alpha - Q_{X_1^T}(x_1^T))_+ \ge \epsilon$*,*

$$\beta_1^*(Q_{X_1^T}, \alpha, \epsilon) = \Big(\sum\nolimits_{x_1^T} (Q_{X_1^T}(x_1^T) - \alpha)_+ - \epsilon \Big)_+.$$

The proof of deferred to Appendix C. Theorem 1 shows that, for any watermarking scheme, the fundamental limits of detection performance depend on the original NTP distribution of the LLM.

When the original $Q_{X_1^T}$ has lower entropy, the minimum achievable detection error increases. This hints that it is inherently difficult to watermark low-entropy text. However, increasing the allowable distortion ϵ can enhance the capacity for reducing detection errors, as illustrated in Figure 3. Moreover, we find that $\beta_1^*(Q_{X_1^T}, \alpha, \epsilon)$ matches the minimum Type-II error from Huang et al. (2023, Theorem 3.2), which is notably optimal for their specific detector. Our results, however, establish that this is the universally minimum Type-II error across all possible detectors and watermarking schemes, indicating that their detector belongs to the set of optimal detectors described below.

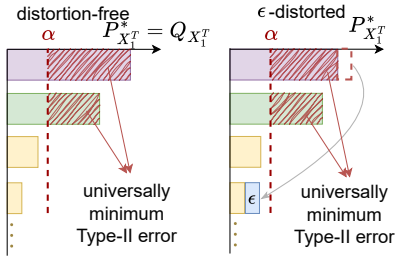

Figure 3: Universally minimum Type-II error w/o distortion

Jointly Optimal Design. We now present the jointly optimal watermarking schemes and detectors that achieve the universally minimum Type-II error in Theorem 1, i.e., the solution to (Opt-O).

Theorem 2 ((Informal Statement) Jointly optimal type of watermarking schemes and detectors)**.** *For any* $(Q_{X_1^T}, \epsilon)$*, the class of optimal detectors is given by*

$$\gamma^* \in \Gamma^* := \big\{ \gamma \mid \gamma(X_1^T, \zeta_1^T) = \mathbb{1}\{X_1^T = g(\zeta_1^T)\}, \text{ for some surjective } g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T \big\}. \tag{4}$$

The corresponding optimal ϵ*-distorted watermarking scheme* $P^*_{X_1^T,\zeta_1^T}$*, whose marginal distribution on* X_1^T *is* $P^*_{X_1^T}$ *(c.f.* (3)*), depends on the original distribution* $Q_{X_1^T}$ *and is detailed in Appendix E.*

Notably, the class of optimal detectors Γ^* are universally optimal. This means that to guarantee the construction of a watermarking scheme that maximizes the detection performance, we must choose a detector from the class Γ^*. Detailed proofs are provided in Appendix E.

Discussions on Theoretically Optimal Watermarking Scheme. For any optimal detector $\gamma^* \in \Gamma^*$ characterized by some function g, constructing the corresponding optimal watermarking scheme

$P^*_{X_1^T,\zeta_1^T}$ is equivalent to transporting the probability mass $P^*_{X_1^T}$ on $\mathcal{V}^T$ to $\mathcal{Z}^T$, making $P^*_{\zeta_1^T|X_1^T}(\zeta_1^T|x_1^T)$ nearly deterministic for $\gamma^*(x_1^T,\zeta_1^T)=1$, while keeping the worst-case Type-I error below α. A detailed illustration is provided in Appendix F, with several important remarks as follows.

First, we observe that the derived optimal watermarking scheme $P^*_{X_1^T,\zeta_1^T}$ for any $\gamma^*\in\Gamma^*$ is *adaptive* to the original LLM output distribution $Q_{X_1^T}$. This observation suggests that, to maximize watermark detection performance, watermarking schemes should fully leverage generative modeling and make the sampling of auxiliary sequence adaptive to $Q_{X_1^T}$. This approach contrasts with existing watermarking schemes, which typically sample the auxiliary sequence according to a given uniform distribution, without adapting it to the LLM NTP distribution. This critical insight serves as a foundation for the design of a practical watermarking scheme, which will be introduced in Section 4.

Second, in order to control the worst-case Type-I error, the construction of $P^*_{X_1^T,\zeta_1^T}$ enlarges the auxiliary sequence set $\mathcal{Z}^T$ by including a redundant sequence $\tilde{\zeta}_1^T$ such that $\gamma(x_1^T,\tilde{\zeta}_1^T)=0$ for all x_1^T. This redundant auxiliary sequence $\tilde{\zeta}_1^T$ plays a critical role in our proposed algorithm.

Third, the optimal watermarking scheme $P^*_{X_1^T,\zeta_1^T}$ is particularly effective in reducing the false alarm rate for low-entropy texts. Specifically, if $P^*_{X_1^T}(x_1^T)>\alpha$ (indicating low-entropy), the text may be mapped to a redundant auxiliary sequence, making it harder to detect as watermarked-LLM-generated.

Lastly, we highlight that our framework and optimal results can be extended to encompass scenarios involving a wide range of attacks, including *semantic-invariant attacks*. In Appendix L, we establish the theoretical foundations for optimal *robust* watermarking schemes and detectors. These findings offer valuable insights for designing advanced semantic-based watermarking algorithms that are resilient to such attacks in the future.

Practical Challenges. While we have derived the theoretically optimal structure, there are still a few practical challenges in its direct implementation. ① Designing a proper function g, an alphabet $\mathcal{Z}^T$ and the corresponding $P^*_{X_1^T,\zeta_1^T}$ is challenging, as $|\mathcal{V}|^T$ grows exponentially with T, making it hard to identify all pairs (x_1^T,ζ_1^T) such that $x_1^T=g(\zeta_1^T)$. ② The optimality is derived for static scenarios with a fixed token length T, making it unsuitable for dynamic scenarios where the tokens are generated incrementally with varying T. ③ In the theoretical analysis, we assume full recovery of the auxiliary sequence ζ_1^T during detection. However, in practice, the detector only receives the token sequence X_1^T, and reconstructing the auxiliary sequence ζ_1^T from X_1^T poses a challenge.

These practical constraints motivate the development of a more feasible version of the theoretically optimal scheme. In Section 4, we extend it to a practical token-level optimal scheme to address ① and ②; in Section 5, we propose an algorithm utilizing a surrogate language model and the Gumbel-max trick Gumbel (1954) to overcome ③.

4 Practical Token-level Optimal Watermarking Scheme

In this section, we present a practical approach that approximates the theoretical framework while ensuring its applicability to real-world scenarios. Building on the fixed-length optimal scheme, we naturally extend it to accommodate varying-length scenarios by constructing the optimal watermarking scheme incrementally for each token, i.e., $P^*_{X_t,\zeta_t|x_1^{t-1},\zeta_1^{t-1}}$ for all $t=1,2,\ldots$.

To lay the groundwork, we first revisit several heuristic detectors for some existing watermarking schemes.

Example 2 (Examples of heuristic detectors)**.** *Two example detectors from existing works:*

- *Green-Red List watermark detector (Kirchenbauer et al., 2023a):* $\gamma(X_1^T,\zeta_1^T)=\mathbb{1}\{\frac{2}{\sqrt{T}}(\sum_{t=1}^T\mathbb{1}\{\zeta_t(X_t)=1\}-\rho T)\geq\lambda\}$ *where* $\lambda>0$*,* $\rho\in(0,1)$*, and* $\zeta_t=(\zeta_t(x))_{x\in\mathcal{V}}$ *is uniformly sampled from* $\{\zeta\in\{0,1\}^{|\mathcal{V}|}:\|\zeta\|_1=\rho|\mathcal{V}|\}$ *with the seed* $\mathrm{hash}(X_{t-1},\mathtt{key})$.
- *Gumbel-max watermark detector (Aaronson, 2023):* $\gamma(X_1^T,\zeta_1^T)=\mathbb{1}\{-\sum_{t=1}^T\log(1-\zeta_t(X_t))\geq\lambda\}\}$ *where* $\lambda>0$*, and* $\zeta_t=(\zeta_t(x))_{x\in\mathcal{V}}$ *is uniformly sampled from* $[0,1]^{|\mathcal{V}|}$ *with the seed* $\mathrm{hash}(X_{t-1}^{t-n},\mathtt{key})$ *for some* n*.*

Practical Detector Design. We observe that the commonly used heuristic detectors take the non-optimal form by averaging the test statistics over each token: $\gamma(X_1^T,\zeta_1^T)=$

$\mathbb{1}\{\frac{1}{T}\sum_{t=1}^{T}$ Test Statistics of $(X_t, \zeta_t) \geq \lambda\}$. This token-level design provides several advantages: (1) incremental computation of detectors for any T and 2) token-level watermarking with the alphabet depending only on the fixed size $|\mathcal{V}|$. Inspired by these detectors, we propose the following detector to address the issues ① and ② mentioned earlier:

$$\gamma_{\mathrm{tk}}(X_1^T, \zeta_1^T) = \mathbb{1}\left\{\frac{1}{T}\sum_{t=1}^{T} \underbrace{\mathbb{1}\{X_t = g_{\mathrm{tk}}(\zeta_t)\}}_{\text{Token-level adaptation of (4)}} \geq \lambda\right\}, \tag{5}$$

for some surjective function $g_{\mathrm{tk}} : \mathcal{Z} \to \mathcal{S} \supset \mathcal{V}$. This detector combines the advantages of existing token-level detectors with the optimal design from Theorem 2. The test statistic for each token (X_t, ζ_t) is optimal at position t, enabling a token-level optimal watermarking scheme that improves the detection performance for each token.

Token-Level Optimal Watermarking Scheme. Following the same rule in Theorem 2 and Appendix E, the token-level optimal watermarking scheme is *sequentially* constructed based on $\mathbb{1}\{X_t = g_{\mathrm{tk}}(\zeta_t)\}$ in (5) and the NTP distribution at each position t, acting only on the token vocabulary $\mathcal{V}$. This approach addresses the challenges ① and ② as well. Notably, the resulting distribution of the token-level optimal scheme for the auxiliary variable ζ_t is adaptive to the original NTP distribution $Q_{X_t|x_1^{t-1}}$. Moreover, the resulting distribution on X_t is given by (comparable to $P^*_{X_1^T}$ in Theorem 2)

$$P^*_{X_t|x_1^{t-1}} := \underset{P_{X_t|x_1^{t-1}} : \mathsf{D}(P_{X_t|x_1^{t-1}}, Q_{X_t|x_1^{t-1}}) \leq \epsilon}{\arg\min} \sum_{x\in\mathcal{V}} (P_{X_t|x_1^{t-1}}(x) - \eta)_+, \tag{6}$$

where $\eta \in (0, 1)$ is the *token-level false alarm constraint*, which is typically much greater than the sequence-level false alarm constraint α. With a proper choice of η, we can effectively control α. Under this scheme, we add watermarks to the generated tokens incrementally, with maximum detection performance at each token. The details are deferred to Appendix G and the algorithm is provided in Section 5.

Performance Analysis. We evaluate the Type-I and Type-II errors of this scheme over the entire sequence (cf. (1)).

Lemma 3 ((Informal Statement) Token-level optimal watermarking detection errors)**.** *Under the detector γ_{tk} in* (5) *and its corresponding token-level optimal watermarking scheme with $\eta \in (0, \min\{1, (\alpha/\binom{T}{\lceil T\lambda\rceil})^{\frac{1}{\lceil T\lambda\rceil}}\})$, the worst-case Type-I error for a length-T sequence is upper bounded by α. If we assume that two tokens with a positional distance greater than n are independent, with a proper detector threshold, the Type-II error decays exponentially in $\frac{T}{n}$.*

Although the token-level optimal watermarking scheme may not be optimal on the entire token sequence, we show that it maintains good performance with a proper choice of token-level false alarm rate η. The formal statement is provided in Appendix H.

Furthermore, we observe that even without explicitly introducing robustness to the token-level optimal watermarking scheme, it inherently leads to some robustness against token replacement. The following result shows that if the auxiliary sequence ζ_1^T is shared between the LLM and the detector γ_{tk} (cf. (5)), the token at position t can be replaced with probability $\Pr(\zeta_t \text{ is redundant})$ without affecting detector output.

Proposition 4 (Robustness against token replacement)**.** *Under the detector γ_{tk} in* (5) *and its corresponding token-level optimal watermarking scheme, the expected number of tokens that can be randomly replaced in X_1^T without compromising detection performance is $\sum_{t=1}^{T} \mathbb{E}_{X_1^{t-1}}\left[\sum_{x\in\mathcal{V}} \left(P^*_{X_t|X_1^{t-1}}(x|X_1^{t-1}) - \eta\right)_+\right]$, with $P^*_{X_t|X_1^{t-1}}$ given in* (6)*.*

5 DAWA: DISTRIBUTION-ADAPTIVE WATERMARKING ALGORITHM

In this section, we address the challenge ③ of transmitting the auxiliary sequence to the detector without knowledge of the original LLM and prompt, using some novel tricks. Building on our proposed token-level optimal watermarking scheme and these innovations, we develop the DAWA (**D**istribution-**A**daptive **W**atermarking **A**lgorithm).

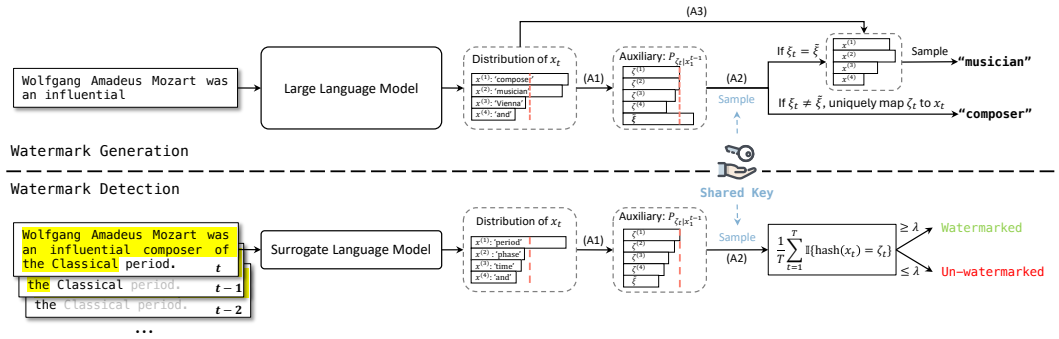

Figure 4: Workflow of our practical algorithm (DAWA) for watermark generation and detection. (A1): construct the distribution of auxiliary variable ζ_t based on the NTP distribution of x_t; (A2): sample ζ_t using the Gumbel-max trick and a shared key; (A3): adjust the NTP distribution of x_t with η.

5.1 Novel Trick for Auxiliary Sequence Transmission

Since the resulting optimal distribution of the auxiliary variable ζ_t is adaptive to the original NTP distribution of LLM, it is not likely to completely reconstruct it at the detection phase without knowledge of the LLM. One possible workaround is enforcing $P_{\zeta_t} = \text{Unif}(\mathcal{Z})$ for both watermark generation and detection. While this method (cf. Appendix I) simplifies the transmission, it leads to a much higher minimum Type-II error compared to $\beta_1^*(Q_{X_1^T}, \alpha, \epsilon)$ (cf. (2)), indicating a trade-off between detection performance and a non-distribution-adaptive design.

We thus introduce a novel trick to transmit the auxiliary sequence by integrating a *surrogate language model (SLM)* during the detection phase and the Gumbel-max trick Gumbel (1954) for sampling ζ_t. This SLM, much smaller than the watermarked LLMs but with the same tokenizer, approximates the watermarked distributions $\{P^*_{X_t|X_1^{t-1}}\}_{t=1,2,\ldots}$ using only the text X_1^T, without the prompt. With the approximated $P^*_{X_t|X_1^{t-1}}$, we reconstruct the sampling distribution of ζ_t and sample it using the Gumbel-max trick with the `key` shared from watermark generation.

In Section 6, our experiments highlight that, even with incomplete recovery of ζ_1^T during detection, the DAWA algorithm with this novel trick exhibits superior detection performance and greater resilience against token replacement attack, surpassing baseline watermarking schemes.

5.2 DAWA

We first design an efficient and practical detector (cf. (5)) by defining g_{tk} as the inverse of a hash function $h_{\texttt{key}}$:

$$\gamma_{\text{dawa}}(X_1^T, \zeta_1^T) = \mathbb{1}\Big\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}\{h_{\texttt{key}}(X_t) = \zeta_t\} \geq \lambda\Big\}. \tag{7}$$

The DAWA is developed on the distortion-free token-level optimal watermarking scheme with $\epsilon = 0$, sampling the auxiliary variable adaptively based on the LLM NTP distribution, as illustrated in Figure 4 and detailed in Appendix J. Below, we elaborate on the key steps.

Watermarked Text Generation. Using the detector γ_{dawa} from (7), we define the auxiliary alphabet $\mathcal{Z}$ from unique mappings $\{h_{\texttt{key}}(x)\}_{x\in\mathcal{V}}$ plus a redundant $\tilde{\zeta}$. At each t, $P_{\zeta_t|x_1^{t-1},\text{pt}}$ is adaptive to $Q_{X_t|x_1^{t-1},\text{pt}}$:

$$\begin{cases} P_{\zeta_t|x_1^{t-1},\text{pt}}(\zeta) \leftarrow (Q_{X_t|x_1^{t-1},\text{pt}}(h^{-1}_{\texttt{key}}(\zeta)) \wedge \eta), \forall \zeta \in \mathcal{Z}\backslash\{\tilde{\zeta}\}. \\ P_{\zeta_t|x_1^{t-1},\text{pt}}(\tilde{\zeta}) \leftarrow \sum_{x\in\mathcal{V}}(Q_{X_t|x_1^{t-1},\text{pt}}(x) - \eta)_+. \end{cases} \tag{A1}$$

The Gumbel-max trick is then used to sample ζ_t:

$$\zeta_t \leftarrow \arg\max_{\zeta\in\mathcal{Z}} \log(P_{\zeta_t|x_1^{t-1},\text{pt}}(\zeta)) + G_{t,\zeta}. \tag{A2}$$

where $G_{t,\zeta}$ is sampled from the Gumbel distribution using a shared `key` and the previous tokens. If ζ_t is non-redundant, $x_t = h^{-1}_{\texttt{key}}(\zeta_t)$; otherwise, x_t is sampled via a multinomial distribution:

$$x_t \sim \left(\frac{(Q_{X_t|x_1^{t-1},\text{pt}}(x) - \eta)_+}{\sum_{x\in\mathcal{V}}(Q_{X_t|x_1^{t-1},\text{pt}}(x) - \eta)_+}\right)_{x\in\mathcal{V}}. \tag{A3}$$

Table 1: Detection performance across different LLMs and datasets.

LLMs	Methods	C4			ELI5		
		ROC-AUC	TP@1% FP	TP@10% FP	ROC-AUC	TP@1% FP	TP@10% FP
Llama2-13B	KGW-1	0.995	0.991	1.000	0.989	0.974	0.986
	EXP-edit	0.986	0.968	0.996	0.983	0.960	0.995
	Gumbel-Max	0.996	0.993	0.994	0.999	0.991	0.994
	Ours	0.999	0.998	1.000	0.998	0.997	1.000
Mistral-8×7B	KGW-1	0.997	0.995	1.000	0.993	0.983	0.994
	EXP-edit	0.993	0.970	0.997	0.994	0.972	0.996
	Gumbel-Max	0.994	0.989	0.999	0.987	0.970	0.990
	Ours	0.999	0.998	1.000	0.999	0.999	1.000

Table 2: Detection performance under token replacement attack.

LLMs	Methods	C4			ELI5		
		ROC-AUC	TP@1% FP	TP@10% FP	ROC-AUC	TP@1% FP	TP@10% FP
Llama2-13B	KGW-1	0.965	0.833	0.952	0.973	0.892	0.973
	EXP-edit	0.973	0.857	0.978	0.967	0.889	0.975
	Gumbel-Max	0.968	0.858	0.970	0.965	0.887	0.975
	Ours	0.989	0.860	0.976	0.995	0.969	0.994
Mistral-8×7B	KGW-1	0.977	0.860	0.962	0.969	0.890	0.970
	EXP-edit	0.980	0.861	0.975	0.983	0.932	0.988
	Gumbel-Max	0.972	0.865	0.960	0.971	0.889	0.975
	Ours	0.990	0.881	0.966	0.993	0.991	0.995

Watermarked Text Detection. A surrogate NTP distribution $\tilde{Q}_{X_t|x_1^{t-1}}$ is approximated by the SLM for each t. We then use (A1) to approximate $P_{\zeta_t|x_1^{t-1},\mathrm{pt}}$ from $\tilde{Q}_{X_t|x_1^{t-1}}$ and sample ζ_t using (A2) with the shared `key`. At each position t, the score $\mathbb{1}\{h_{\texttt{key}}(x_t) = \zeta_t\}$ is 1 if ζ_t non-redundant and 0 otherwise. Compute $\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{h_{\texttt{key}}(x_t) = \zeta_t\}$ and compare with a threshold λ. If above λ, the text is detected as watermarked.

6 EXPERIMENTS

The specifics of our experimental setup are provided in Appendix A, covering implementation details, baselines (KGW-1 (Kirchenbauer et al., 2023a), EXP-edit (Kuditipudi et al., 2023), and Gumbel-Max (Aaronson, 2023)), datasets (C4 (Raffel et al., 2020a) and ELI5 (Fan et al., 2019)) and prompts, as well as evaluation metrics (ROC-AUC score, True Positive (TP) Rate and False Positive (FP) Rate).

6.1 MAIN EXPERIMENTAL RESULTS

Watermark Detection Performance. A low FPR is essential to avoid incorrectly identifying unwatermarked text as watermarked. To explore our detection performance at a very low FPR, we conduct experiments using Llama2-13B on 100k texts from the Wikipedia dataset and compute the TPR at $1\mathrm{e}{-01}$, $1\mathrm{e}{-02}$, $1\mathrm{e}{-03}$, $1\mathrm{e}{-04}$, and $1\mathrm{e}{-05}$ FPR respectively. Figure 1 shows that DAWA significantly outperforms other baselines. Furthermore, we compare the detection performance across various language models and tasks, as presented in Table 1. Our watermarking method demonstrates superior performance, especially on the relatively low-entropy QA dataset. This success stems from the design of our watermarking scheme, which reduces the likelihood of low-entropy tokens being falsely detected as watermarked, thereby lowering the FPR. Moreover, this suggests that even without knowing the watermarked LLM during detection, we can still use the proposed SLM and Gumbel-max trick to successfully detect the watermark.

Robustness. We assess the robustness of our watermarking methods against a token replacement attack. As discussed in Proposition 4, the proposed token-level optimal watermarking scheme has inherent robustness against token replacement. For each watermarked text, we randomly mask 50% of the tokens and use T5-large (Raffel et al., 2020b) to predict the replacement for each masked token based on the context. For each prediction, the predicted token retains a chance of being the original one, as we do not force the replacement to differ from the original to maintain the sentence's semantics and quality. Yet, about 35% of tokens in watermarked sentences are still replaced on average. Table 2 exhibits watermark detection performance under token replacement attacks across different language models and tasks. It presents the robustness of our proposed watermarking method against the token replacement attack. Our method remains high ROC-AUC, TPR@1%FPR, and TPR@10%FPR under this attack compared with other baselines. This result supports our theoretical analysis on robustness in Proposition 4.

Watermarked Text Quality. To evaluate the quality of watermarked text generated by our watermarking methods, we report the perplexity on C4 dataset using GPT-3 (Brown et al., 2020), and the BLEU score on the machine translation task using the WMT19 dataset (Barrault et al., 2019) and MBART Model (Liu et al., 2020), as shown in Table 3. It can be observed that our scheme achieves a higher BLEU score and a lower perplexity than the baseline distortion-free schemes and is close to the score on datasets. This demonstrates that our scheme, employing an NTP distribution-adaptive approach, has minimal impact on the generated text quality, preserving its naturalness and coherence.

Table 3: Comparison of BLEU score and average perplexity across different watermarking methods.

Methods	Human	KGW-1	EXP-Edit	Gumbel-Max	**Ours**
BLEU Score	0.219	0.158	0.203	0.210	0.214
Avg Perplexity	8.846	14.327	12.186	11.732	6.495

Additional Results. In Appendix K, we show that: (1) our theoretical choice of η effectively controls the empirical FPR; (2) our watermarking scheme does not affect generation time; and (3) detection remains accurate and robust even with a much smaller SLM from a different model family and without prompts.

REFERENCES

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

Sahar Abdelnabi and Mario Fritz. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 121–140. IEEE, 2021.

Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Marcis Fishel, Yvette Graham, Francisco Guzmán, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Aurélie Névéol, Günter Neumann, Adrià de Gispert Pastor, Matt Post, Raphaël Rubino, Lucia Specia, and Marcos Zampieri. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 1–61, 2019.

Massieh Kordi Boroujeny, Ya Jiang, Kai Zeng, and Brian Mark. Multi-bit distortion-free watermarking for large language models. *arXiv preprint arXiv:2402.16578*, 2024.

Jack T Brassil, Steven Low, Nicholas F Maxemchuk, and Lawrence O'Gorman. Electronic marking and identification techniques to discourage document copying. *IEEE Journal on Selected Areas in Communications*, 13(8):1495–1504, 1995.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 1877–1901, 2020.

Brian Chen. *Design and analysis of digital watermarking, information embedding, and data hiding systems*. PhD thesis, Massachusetts Institute of Technology, 2000.

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

Jaiden Fairoze, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, and Mingyuan Wang. Publicly detectable watermarking for language models. *arXiv preprint arXiv:2310.18491*, 2023.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*, 2019.

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. Three bricks to consolidate watermarks for large language models. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6. IEEE, 2023.

Jiayi Fu, Xuandong Zhao, Ruihan Yang, Yuansen Zhang, Jiangjie Chen, and Yanghua Xiao. Gumbelsoft: Diversified language model watermarking via the GumbelMax-trick. *arXiv preprint arXiv:2402.12948*, 2024.

Eva Giboulot and Teddy Furon. Watermax: breaking the LLM watermark detectability-robustness-quality trade-off. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=HjeKHxK2VH.

E.J. Gumbel. *Statistical Theory of Extreme Values and Some Practical Applications: A Series of Lectures*. Applied mathematics series. U.S. Government Printing Office, 1954. URL https://books.google.com/books?id=SNpJAAAAMAAJ.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. *arXiv preprint arXiv:2310.10669*, 2023.

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

Mingjia Huo, Sai Ashish Somayajula, Youwei Liang, Ruisi Zhang, Farinaz Koushanfar, and Pengtao Xie. Token-specific watermarking with enhanced detectability and semantic coherence for large language models. *arXiv preprint arXiv:2402.18059*, 2024.

Svante Janson. New versions of suen's correlation inequality. *Random Structures and Algorithms*, 13 (3-4):467–483, 1998.

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

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*, pp. 17061–17084. PMLR, 2023a.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. *arXiv preprint arXiv:2306.04634*, 2023b.

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

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, 2024.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A semantic invariant robust watermark for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Yepeng Liu and Yuheng Bu. Adaptive text watermark for large language models. In *Forty-first International Conference on Machine Learning*, 2024.

Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

Emin Martinian, Gregory W Wornell, and Brian Chen. Authentication with distortion criteria. *IEEE Transactions on Information Theory*, 51(7):2523–2542, 2005.

N. Merhav and E. Ordentlich. On causal and semicausal codes for joint information embedding and source coding. *IEEE Transactions on Information Theory*, 52(1):213–226, 2006. doi: 10.1109/TIT.2005.860428.

Neri Merhav and Erez Sabbag. Optimal watermark embedding and detection strategies under limited detection resources. *IEEE Transactions on Information Theory*, 54(1):255–274, 2008. doi: 10.1109/TIT.2007.911210.

Pierre Moulin and Joseph A O'Sullivan. Information-theoretic analysis of watermarking. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 6, pp. 3630–3633. IEEE, 2000.

Travis Munyer, Abdullah Tanvir, Arjon Das, and Xin Zhong. Deeptextmark: A deep learning-driven text watermarking approach for identifying large language model generated text. *arXiv preprint arXiv:2305.05773*, 2023.

Wenjie Qu, Dong Yin, Zixin He, Wei Zou, Tianyang Tao, Jinyuan Jia, and Jiaheng Zhang. Provably robust multi-bit watermarking for AI-generated text via error correction code. *arXiv preprint arXiv:2401.16820*, 2024.

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

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, 2020a.

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, 2020b.

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

Ryoma Sato, Yuki Takezawa, Han Bao, Kenta Niwa, and Makoto Yamada. Embarrassingly simple text watermarks. *arXiv preprint arXiv:2310.08920*, 2023.

Hermann Thorisson. Coupling methods in probability theory. *Scandinavian journal of statistics*, pp. 159–182, 1995.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Shangqing Tu, Yuliang Sun, Yushi Bai, Jifan Yu, Lei Hou, and Juanzi Li. Waterbench: Towards holistic evaluation of watermarks for large language models. *arXiv preprint arXiv:2311.07138*, 2023.

Lean Wang, Wenkai Yang, Deli Chen, Hao Zhou, Yankai Lin, Fandong Meng, Jie Zhou, and Xu Sun. Towards codable text watermarking for large language models. *arXiv preprint arXiv:2307.15992*, 2023.

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

Xi Yang, Jie Zhang, Kejiang Chen, Weiming Zhang, Zehua Ma, Feng Wang, and Nenghai Yu. Tracing text provenance via context-aware lexical substitution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11613–11621, 2022.

Xi Yang, Kejiang Chen, Weiming Zhang, Chang Liu, Yuang Qi, Jie Zhang, Han Fang, and Nenghai Yu. Watermarking text generated by black-box language models. *arXiv preprint arXiv:2305.08883*, 2023.

KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. Robust multi-bit natural language watermarking through invariant features. *arXiv preprint arXiv:2305.01904*, 2023.

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

Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Farinaz Koushanfar. {REMARK-LLM}: A robust and efficient watermarking framework for generative large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 1813–1830, 2024.

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

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

Xuandong Zhao, Lei Li, and Yu-Xiang Wang. Permute-and-flip: An optimally robust and watermarkable decoder for LLMs. *arXiv preprint arXiv:2402.05864*, 2024.

Tong Zhou, Xuandong Zhao, Xiaolin Xu, and Shaolei Ren. Bileve: Securing text provenance in large language models against spoofing with bi-level signature. *arXiv preprint arXiv:2406.01946*, 2024.

A EXPERIMENT SETTINGS

Implementation Details. Our approach is implemented on two language models: Llama2-13B (Touvron et al., 2023), and Mistral-8×7B (Jiang et al., 2023). Llama2-7B serves as the surrogate model for Llama2-13B, while Mistral-7B is used as the surrogate model for Mistral-8×7B. We conduct our experiments on Nvidia A100 GPUs. In the DAWA, we set $\eta = 0.2$ and $T = 200$.

Baselines. We compare our methods with three existing watermarking methods: KGW-1 (Kirchenbauer et al., 2023a), EXP-edit (Kuditipudi et al., 2023), and Gumbel-Max (Aaronson, 2023), where the EXP-edit and Gumbel-Max are distortion-free watermark. KGW-1 employs the prior 1 token as a hash to create a green/red list, with the watermark strength set at 2.

Dataset and Prompt. Our experiments are conducted using two distinct datasets. The first is an open-ended **high-entropy** generation dataset, a realnewslike subset from C4 (Raffel et al., 2020a). The second is a relatively **low-entropy** generation dataset, ELI5 (Fan et al., 2019). The realnewslike subset of C4 is tailored specifically to include high-quality journalistic content that mimics the style and format of real-world news articles. We utilize the first two sentences of each text as prompts and the following 200 tokens as human-generated text. The ELI5 dataset is specifically designed for the task of long-form question answering (QA), with the goal of providing detailed explanations for complex questions. We use each question as a prompt and its answer as human-generated text.

Evaluation Metrics. To evaluate the performance of watermark detection, we report the ROC-AUC score, where the ROC curve shows the True Positive (TP) Rate against the False Positive (FP) Rate. A higher ROC-AUC score indicates better overall performance. Additionally, we report the TP at FR values ranging from $1\mathrm{e}{-01}$ to $1\mathrm{e}{-05}$ to specifically assess detection accuracy with a particularly low risk of falsely classifying unwatermarked text as watermarked. The detection threshold λ is determined empirically by the ROC-AUC score function based on unwatermarked and watermarked sentences.

B OTHER EXISTING WATERMARKING SCHEMES

Here, we discuss additional existing watermarking schemes utilizing auxiliary variables, which can be encompassed within our LLM watermarking formulation.

• The **Gumbel-max watermarking scheme** (Aaronson, 2023) applies the Gumbel-max trick (Gumbel, 1954) to sample the next token X_t, where the Gumbel variable is exactly the auxiliary variable ζ_t, which is a $|\mathcal{V}|$-dimensional vector, indexd by x. For $t = 1, 2, \ldots,$

– Compute a hash using the previous n tokens X_{t-1}^{t-n} and a shared secret `key`, i.e., $h(X_{t-1}^{t-n}, \texttt{key})$, where $h : \mathcal{V}^n \times \mathbb{R} \to \mathbb{R}$.
– Use $h(X_{t-1}^{t-n}, \texttt{key})$ as a seed to uniformly sample the auxiliary vector ζ_t from $[0, 1]^{|\mathcal{V}|}$.
– Sample X_t using the Gumbel-max trick

$$X_t = \arg\max_{x \in \mathcal{V}} \log Q_{X_t|x_1^{t-1}}(x) - \log(-\log \zeta_t(x)).$$

• In the **inverse transform watermarking scheme** (Kuditipudi et al., 2023), the vocabulary $\mathcal{V}$ is considered as $[|\mathcal{V}|]$ and the combination of the uniform random variable and the randomly permuted index vector is the auxiliary variable ζ_t.

– Use `key` as a seed to uniformly and independently sample $\{U_t\}_{t=1}^T$ from $[0, 1]$, and $\{\pi_t\}_{t=1}^T$ from the space of permutations over $[|\mathcal{V}|]$. Let the auxiliary variable $\zeta_t = (U_t, \pi_t)$, for $t = 1, 2, \ldots, T$.
– Sample X_t as follows

$$X_t = \pi_t^{-1}\left(\min\left\{ i \in [|\mathcal{V}|] : \sum_{x \in [|\mathcal{V}|]} \left(Q_{X_t|x_1^{t-1}}(x) \mathbb{1}\{\pi_t(x) \le i\} \right) \ge U_t \right\} \right),$$

where π_t^{-1} denotes the inverse permutation.

• In **adaptive watermarking** by Liu & Bu (2024), the authors introduce a watermarking scheme that adopts a technique similar to the Green-Red List approach but replaces the hash function with a pretrained neural network h. The auxiliary variable ζ_t is sampled from the set $\{\mathbf{v} \in \{0, 1\}^{|\mathcal{V}|} : \|\mathbf{v}\|_1 = \rho|\mathcal{V}|\}$ using the seed $h(\phi(X_1^{t-1}), \texttt{key})$, where h takes the semantics $\phi(X_1^{t-1})$ of the

generated text and the secret `key` as inputs. They sample X_t using the same process as the Green-Red List approach.

C PROOF OF THEOREM 1

According to the Type-I error constraint, we have $\forall x_1^T \in \mathcal{V}^T$,

$$\begin{aligned}
\alpha &\geq \max_{Q_{X_1^T}} \mathbb{E}_{Q_{X_1^T} P_{\zeta_1^T}}[\mathbb{1}\{(X_1^T, \zeta_1^T) \in \mathcal{A}_1\}] \\
&\geq \mathbb{E}_{\delta_{x_1^T} P_{\zeta_1^T}}[\mathbb{1}\{(X_1^T, \zeta_1^T) \in \mathcal{A}_1\}] \\
&= \mathbb{E}_{P_{\zeta_1^T}}[\mathbb{1}\{(x_1^T, \zeta_1^T) \in \mathcal{A}_1\}] \\
&= \begin{cases} \sum_{\zeta_1^T} P_{\zeta_1^T}(\zeta_1^T) \mathbb{1}\{(x_1^T, \zeta_1^T) \in \mathcal{A}_1\}, & \mathcal{Z} \text{ is discrete;} \\ \int P_{\zeta_1^T}(\zeta_1^T) \mathbb{1}\{(x_1^T, \zeta_1^T) \in \mathcal{A}_1\} \, d\zeta_1^T, & \mathcal{Z} \text{ is continuouts;} \end{cases}.
\end{aligned}$$

In the following, for notational simplicity, we assume that $\mathcal{Z}$ is discrete. However, the derivations hold for both discrete $\mathcal{Z}$ and continuous $\mathcal{Z}$. The Type-II error is given by $1 - \mathbb{E}_{P_{X_1^T, \zeta_1^T}}[\mathbb{1}\{(X_1^T, \zeta_1^T) \in \mathcal{A}_1\}]$. We have

$$\mathbb{E}_{P_{X_1^T, \zeta_1^T}}[\mathbb{1}\{(X_1^T, \zeta_1^T) \in \mathcal{A}_1\}] = \sum_{x_1^T} \underbrace{\sum_{\zeta_1^T} P_{X_1^T, \zeta_1^T}(x_1^T, \zeta_1^T) \mathbb{1}\{(x_1^T, \zeta_1^T) \in \mathcal{A}_1\}}_{C(x_1^T)}, \tag{8}$$

where for all $x_1^T \in \mathcal{V}^T$,

$$C(x_1^T) \leq P_{X_1^T}(x_1^T) \quad \text{and} \quad C(x_1^T) \leq \sum_{\zeta_1^T} P_{\zeta_1^T}(\zeta_1^T) \mathbb{1}\{(x_1^T, \zeta_1^T) \in \mathcal{A}_1\} \leq \alpha$$

according to the Type-I error bound. Therefore,

$$\begin{aligned}
\mathbb{E}_{P_{X_1^T, \zeta_1^T}}[\mathbb{1}\{(X_1^T, \zeta_1^T) \in \mathcal{A}_1\}] &= \sum_{x_1^T} C(x_1^T) \leq \sum_{x_1^T} (P_{X_1^T}(x_1^T) \wedge \alpha) \\
&= 1 - \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+ \qquad (9)
\end{aligned}$$

where (9) is maximized at

$$P_{X_1^T}^* \coloneqq \operatorname*{arg\,min}_{P_{X_1^T} : \mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+. \tag{10}$$

For any $P_{X_1^T}$, the Type-II error is lower bounded by

$$\mathbb{E}_{P_{X_1^T, \zeta_1^T}}[\mathbb{1}\{(X_1^T, \zeta_1^T) \notin \mathcal{A}_1\}] \geq \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+.$$

By plugging $P_{X_1^T}^*$ into this lower bound, we obtain a Type-II lower bound that holds for all γ and $P_{X_1^T, \zeta_1^T}$. Recall that Huang et al. (2023) proposed a type of detector and watermarking scheme that achieved this lower bound. As we demonstrate, it is actually the universal minimum Type-II error over all possible γ and $P_{X_1^T, \zeta_1^T}$, denoted by $\beta_1^*(Q_{X_1^T}, \epsilon, \alpha)$.

Specifically, define $\epsilon^*(x_1^T) = Q_{X_1^T}(x_1^T) - P_{X_1^T}^*(x_1^T)$ and we have

$$\begin{aligned}
\sum_{x_1^T : P_{X_1^T}^*(x_1^T) \geq \alpha} \epsilon^*(x_1^T) &= \sum_{x_1^T : P_{X_1^T}^*(x_1^T) \geq \alpha, \epsilon^*(x_1^T) \geq 0} \epsilon^*(x_1^T) + \underbrace{\sum_{x_1^T : P_{X_1^T}^*(x_1^T) \geq \alpha, \epsilon^*(x_1^T) \leq 0} \epsilon^*(x_1^T)}_{\leq 0} \\
&\leq \sum_{x_1^T : P_{X_1^T}^*(x_1^T) \geq \alpha, \epsilon^*(x_1^T) \geq 0} \epsilon^*(x_1^T)
\end{aligned}$$

$$= \sum_{x_1^T : P^*_{X_1^T}(x_1^T) \geq \alpha, Q_{X_1^T}(x_1^T) \geq P^*_{X_1^T}(x_1^T)} \epsilon^*(x_1^T)$$
$$\leq \sum_{x_1^T : Q_{X_1^T}(x_1^T) \geq P^*_{X_1^T}(x_1^T)} \epsilon^*(x_1^T) \leq \epsilon$$

where the last inequality follows from the total variation distance constraint $\mathsf{D_{TV}}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$. We rewrite $\beta_1^*(Q_{X_1^T}, \epsilon, \alpha)$ as follows:

$$\begin{aligned}
\beta_1^*(Q_{X_1^T}, \epsilon, \alpha) &= \min_{P_{X_1^T} : \mathsf{D_{TV}}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+ \qquad (11)\\
&= \sum_{x_1^T : P^*_{X_1^T}(x_1^T) \geq \alpha} (P^*_{X_1^T}(x_1^T) - \alpha),\\
&= \sum_{x_1^T : P^*_{X_1^T}(x_1^T) \geq \alpha} (Q_{X_1^T}(x_1^T) - \epsilon^*(x_1^T) - \alpha)\\
&= \sum_{x_1^T : P^*_{X_1^T}(x_1^T) \geq \alpha} (Q_{X_1^T}(x_1^T) - \alpha) - \sum_{x_1^T : P^*_{X_1^T}(x_1^T) \geq \alpha} \epsilon^*(x_1^T)\\
&\geq \sum_{x_1^T} (Q_{X_1^T}(x_1^T) - \alpha)_+ - \epsilon,
\end{aligned}$$

where the last inequality follows from $\sum_{x_1^T : P^*_{X_1^T}(x_1^T) \geq \alpha} \epsilon^*(x_1^T) \leq \epsilon$, i.e. the total variation constraint limits how much the distribution $P^*_{X_1^T}$ can be perturbed from $Q_{X_1^T}$. Since $\beta_1^*(Q_{X_1^T}, \epsilon, \alpha) \geq 0$, finally we have

$$\beta_1^*(Q_{X_1^T}, \epsilon, \alpha) \geq \left(\sum_{x_1^T} (Q_{X_1^T}(x_1^T) - \alpha)_+ - \epsilon \right)_+ .$$

Notably, the lower bound is achieved when $\{x_1^T : P^*_{X_1^T}(x_1^T) \geq \alpha\} = \{x_1^T : Q_{X_1^T}(x_1^T) \geq P^*_{X_1^T}(x_1^T)\}$ and $\mathsf{D_{TV}}(Q_{X_1^T}, P^*_{X_1^T}) = \epsilon$. That is, to construct $P^*_{X_1^T}$, an ϵ amount of the mass of $Q_{X_1^T}$ above α is moved to below α, which is possible only when $\sum_{x_1^T} (\alpha - Q_{X_1^T}(x_1^T))_+ \geq \epsilon$. Note that Huang et al. (2023, Theorem 3.2) points out a sufficient condition for this to hold: $|\mathcal{V}|^T \geq \frac{1}{\alpha}$. The optimal distribution $P^*_{X_1^T}$ thus satisfies

$$\sum_{x_1^T : Q_{X_1^T}(x_1^T) \geq \alpha} (Q_{X_1^T}(x_1^T) - P^*_{X_1^T}(x_1^T)) = \sum_{x_1^T : Q_{X_1^T}(x_1^T) \leq \alpha} (P^*_{X_1^T}(x_1^T) - Q_{X_1^T}(x_1^T)) = \epsilon.$$

Refined constraints for optimization. We notice that the feasible region of (Opt-I) can be further reduced as follows:

$$\min_{P_{X_1^T}} \min_{P_{\zeta_1^T | X_1^T}} \quad \mathbb{E}_{P_{X_1^T} P_{\zeta_1^T | X_1^T}}[1 - \gamma(X_1^T, \zeta_1^T)] \qquad \text{(Opt-II)}$$

$$\begin{aligned}
\text{s.t.} \quad & \int P_{\zeta_1^T | X_1^T}(\zeta_1^T | x_1^T)\, d\zeta_1^T = 1, \ \forall x_1^T\\
& \int P_{\zeta_1^T | X_1^T}(\zeta_1^T | x_1^T) \gamma(x_1^T, \zeta_1^T) \leq 1 \wedge \frac{\alpha}{P_{X_1^T}(x_1^T)}, \ \forall x_1^T \qquad (12)\\
& \mathsf{D_{TV}}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon,\\
& \sup_{Q_{X_1^T}} \sum_{x_1^T} Q_{X_1^T}(x_1^T) \int \left(\sum_{y_1^T} P_{\zeta_1^T | X_1^T}(\zeta_1^T | y_1^T) P_{X_1^T}(y_1^T) \right) \gamma(x_1^T, \zeta_1^T)\, d\zeta_1^T \leq \alpha,
\end{aligned}$$

where (12) is an additional constraint on $P_{\zeta_1^T|X_1^T}$. If and only if (12) can be achieved with equality, the minimum of the objective function $\mathbb{E}_{P_{X_1^T}P_{\zeta_1^T|X_1^T}}[1-\gamma(X_1^T,\zeta_1^T)]$ reaches (2).

D An Example of Suboptimal Detectors and its Proof

To better illustrate the universal optimality of the class of detectors Γ^*, we provide an example of suboptimal detectors where *no* watermarking scheme can achieve universally optimal performance.

Example 3 (Suboptimal detectors)**.** *Consider a detector $\gamma(X_1^T,\zeta_1^T) = \mathbb{1}\{f(X_1^T) = \zeta_1^T\}$, for some surjective function $f : \mathcal{V}^T \to \mathcal{S} \subseteq \mathcal{Z}^T$. The minimum Type-II error attained by the corresponding optimal watermarking scheme from* (Opt-I) *is $\min_{P_{X_1^T}:\mathsf{D}(P_{X_1^T},Q_{X_1^T})\le\epsilon} \sum_{s\in\mathcal{S}} \left(\left(\sum_{x_1^T:f(x_1^T)=s} P_{X_1^T}(x_1^T)\right)-\alpha\right)_+$, higher than $\beta_1^*(Q_{X_1^T},\alpha,\epsilon)$.*

In the robustness discussion at the end of the paper, we will further show that this is, in fact, optimal in the presence of certain types of text modifications.

In this proof, we assume that $\mathcal{Z}$ is discrete for simplicity. However, the result holds for continuous $\mathcal{Z}$ without loss of generality. If the detector accepts the form $\gamma(X_1^T,\zeta_1^T) = \mathbb{1}\{f(X_1^T) = \zeta_1^T\}$ for some surjective function $f : \mathcal{V}^T \to \mathcal{S}$ and $\mathcal{S} \subseteq \mathcal{Z}^T$, we have for any $s \in \mathcal{S}$,

$$\begin{aligned}\alpha \ge \sup_{Q_{X_1^T}} \mathbb{E}_{Q_{X_1^T}P_{\zeta_1^T}}[\mathbb{1}\{f(X_1^T)=\zeta_1^T\}] &\ge \mathbb{E}_{P_{\zeta_1^T}}[\mathbb{1}\{s=\zeta_1^T\}]\\ &= \sum_{\zeta_1^T} P_{\zeta_1^T}(\zeta_1^T)\mathbb{1}\{s=\zeta_1^T\},\end{aligned}$$

and (8) can be rewritten as

$$\mathbb{E}_{P_{X_1^T,\zeta_1^T}}[\mathbb{1}\{f(X_1^T)=\zeta_1^T\}] = \sum_{s\in\mathcal{S}} \underbrace{\sum_{x_1^T:f(x_1^T)=s}\sum_{\zeta_1^T} P_{X_1^T,\zeta_1^T}(x_1^T,\zeta_1^T)\mathbb{1}\{f(x_1^T)=\zeta_1^T\}}_{C(s)},$$

where

$$C(s) \le \sum_{x_1^T:f(x_1^T)=s} P_{X_1^T}(x_1^T) \quad \text{and} \quad C(s) \le \sum_{\zeta_1^T} P_{\zeta_1^T}(\zeta_1^T)\mathbb{1}\{s=\zeta_1^T\} \le \alpha.$$

Therefore, the Type-II error for such type of detector γ is lower bounded by

$$\begin{aligned}&\mathbb{E}_{P_{X_1^T,\zeta_1^T}}[\mathbb{1}\{f(X_1^T)\neq\zeta_1^T\}]\\ &= 1-\sum_{s\in\mathcal{S}} C(s) \ge 1-\sum_{s\in\mathcal{S}}\left(\left(\sum_{x_1^T:f(x_1^T)=s} P_{X_1^T}(x_1^T)\right)\wedge\alpha\right)\\ &= \sum_{s\in\mathcal{S}}\left(\left(\sum_{x_1^T:f(x_1^T)=s} P_{X_1^T}(x_1^T)\right)-\alpha\right)_+\\ &\ge \min_{P_{X_1^T}:\mathsf{D}_{\mathsf{TV}}(P_{X_1^T},Q_{X_1^T})\le\epsilon}\sum_{s\in\mathcal{S}}\left(\left(\sum_{x_1^T:f(x_1^T)=s} P_{X_1^T}(x_1^T)\right)-\alpha\right)_+,\end{aligned}$$

where the last inequality holds with equality when

$$P_{X_1^T} = \operatorname*{arg\,min}_{P_{X_1^T}:\mathsf{D}_{\mathsf{TV}}(P_{X_1^T},Q_{X_1^T})\le\epsilon}\sum_{s\in\mathcal{S}}\left(\left(\sum_{x_1^T:f(x_1^T)=s} P_{X_1^T}(x_1^T)\right)-\alpha\right)_+.$$

This minimum achievable Type-II error is higher than $\beta_1^*(Q_{X_1^T},\alpha,\epsilon)$ (cf. (11)) due to the summation over $\{x_1^T : f(x_1^T) = s\}$.

E Formal Statement of Theorem 2 and its Proof

Theorem 2 [Formal] (Optimal type of detectors and watermarking schemes)**.** *The set of all detectors that achieve the minimum Type-II error $\beta_1^*(Q_{X_1^T},\alpha,\epsilon)$ in Theorem 1 for all text distribution $Q_{X_1^T} \in$*

$\mathcal{P}(\mathcal{V}^T)$ and distortion level $\epsilon \geq 0$ is precisely

$$\Gamma^* \coloneqq \big\{\gamma \,|\gamma(X_1^T, \zeta_1^T) = \mathbb{1}\{X_1^T = g(\zeta_1^T)\}, \text{ for some surjective } g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T\big\}.$$

For any valid function g, choose a redundant auxiliary value $\tilde{\zeta_1^T} \in \mathcal{Z}^T$ such that $x_1^T \neq g(\tilde{\zeta_1^T})$ for all $x_1^T \in \mathcal{V}^T$. The detailed construction of the optimal watermarking scheme is as follows:

$$P^*_{X_1^T} = \min_{P_{X_1^T}:\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+,$$

and for any $x_1^T \in \mathcal{V}^T$, $P^*_{\zeta_1^T|X_1^T}(\zeta_1^T|x_1^T)$ satisfies (13)

$$\begin{cases} P^*_{X_1^T}(x_1^T) \sum_{\zeta_1^T} P^*_{\zeta_1^T|X_1^T}(\zeta_1^T|x_1^T)\gamma(x_1^T, \zeta_1^T) = P^*_{X_1^T}(x_1^T) \wedge \alpha, & \forall \zeta_1^T \text{ s.t. } \gamma(x_1^T, \zeta_1^T) = 1; \\ P^*_{X_1^T}(x_1^T) P^*_{\zeta_1^T|X_1^T}(\zeta_1^T|x_1^T) = \big(P^*_{X_1^T}(x_1^T) - \alpha\big)_+, & \text{if } \zeta_1^T = \tilde{\zeta_1^T}; \\ P^*_{\zeta_1^T|X_1^T}(\zeta_1^T|x_1^T) = 0, & \text{otherwise.} \end{cases}$$

Proof. First, we observe that the lower bound on the Type-II error in (2) is attained if and only if the constraint in (12) holds with equality for all x_1^T and for the optimizer. Thus, it suffices to show that for any detector $\gamma \notin \Gamma^*$, the constraint in (12) cannot hold with equality for all x_1^T given any text distributions $Q_{X_1^T}$. First define an arbitrary surjective function $g : \mathcal{Z}^T \to \mathcal{S}$, where $\mathcal{S}$ is on the same metric space as $\mathcal{V}^T$. Cases 1 and 2 prove that $\mathcal{V}^T \subset \mathcal{S}$. Case 3 proves that γ can only be $\gamma(X_1^T, \zeta_1^T) = \mathbb{1}\{X_1^T = g(\zeta_1^T)\}$.

- **Case 1:** $\gamma(X_1^T, \zeta_1^T) = \mathbb{1}\{X_1^T = g(\zeta_1^T)\}$ but $\mathcal{S} \subset \mathcal{V}^T$. There exists $\tilde{x}_1^T$ such that for all ζ_1^T, $\mathbb{1}\{\tilde{x}_1^T = g(\zeta_1^T)\} = 0$. Under this case, (12) cannot hold with equality for $\tilde{x}_1^T$ since the LHS is always 0 while the RHS is positive.

- **Case 2:** $\gamma(X_1^T, \zeta_1^T) = \mathbb{1}\{X_1^T = g(\zeta_1^T)\}$ but $\mathcal{S} = \mathcal{V}^T$. Let us start from the simple case where $T = 1$, $\mathcal{V} = \{x_1, x_2\}$, $\mathcal{Z} = \{\zeta_1, \zeta_2\}$, and g is an identity mapping. Given any Q_X and any feasible P_X such that $\mathsf{D_{TV}}(P_X, Q_X) \leq \epsilon$, when (12) holds with equality, i.e.,

  $$P_{X,\zeta}(x_1, \zeta_1) = P_X(x_1) \wedge \alpha \quad \text{and} \quad P_{X,\zeta}(x_2, \zeta_2) = P_X(x_2) \wedge \alpha,$$

  then the marginal P_ζ is given by: $P_\zeta(\zeta_1) = P_X(x_1) \wedge \alpha + (P_X(x_2) - \alpha)_+$, $P_\zeta(\zeta_2) = P_X(x_2) \wedge \alpha + (P_X(x_1) - \alpha)_+$. The worst-case Type-I error is given by

  $$\sup_{Q_X} \Big(Q_X(x_1)\big(P_X(x_1) \wedge \alpha + (P_X(x_2) - \alpha)_+\big) + Q_X(x_2)\big(P_X(x_2) \wedge \alpha + (P_X(x_1) - \alpha)_+\big) \Big)$$

  $$\geq P_X(x_1) \wedge \alpha + (P_X(x_2) - \alpha)_+$$
  $$> \alpha, \quad \text{if } P_X(x_1) > \alpha, P_X(x_2) > \alpha.$$

  It implies that for any Q_X such that $\{P_X \in \mathcal{P}(\mathcal{V}) : \mathsf{D_{TV}}(P_X, Q_X) \leq \epsilon\} \subseteq \{P_X \in \mathcal{P}(\mathcal{V}) : P_X(x_1) > \alpha, P_X(x_2) > \alpha\}$, the false-alarm constraint is violated when (12) holds with equality. It can be easily verified that this result also holds for larger $(T, \mathcal{V}, \mathcal{Z})$ and other functions $g : \mathcal{Z}^T \to \mathcal{V}^T$.

- **Case 3:** Let $\Xi_\gamma(x_1^T) \coloneqq \{\zeta_1^T \in \mathcal{Z}^T : \gamma(x_1^T, \zeta_1^T) = 1\}$. $\exists x_1^T \neq y_1^T \in \mathcal{V}^T$, s.t. $\Xi(x_1^T) \cap \Xi(y_1^T) \neq \emptyset$. For any detector $\gamma \notin \Gamma^*$ that does not fall into Cases 1 and 2, it falls into Case 3. Let us start from the simple case where $T = 1$, $\mathcal{V} = \{x_1, x_2\}$, $\mathcal{Z} = \{\zeta_1, \zeta_2, \zeta_3\}$. Consider a detector γ as follows: $\gamma(x_1, \zeta_1) = \gamma(x_2, \zeta_1) = 1$ and $\gamma(x, \zeta) = 0$ for all other pairs $(x, \zeta) \in \mathcal{V} \times \mathcal{Z}$. Hence, $\Xi(x_1) \cap \Xi(x_2) = \{\zeta_1\}$. When (12) holds with equality, i.e.,

  $$P_{X,\zeta}(x_1, \zeta_1) = P_X(x_1) \wedge \alpha \quad \text{and} \quad P_{X,\zeta}(x_2, \zeta_1) = P_X(x_2) \wedge \alpha,$$

  we have the worst-case Type-I error lower bounded by

  $$\sup_{Q_X} \Big(Q_X(x_1)P_\zeta(\zeta_1) + Q_X(x_2)P_\zeta(\zeta_1) \Big) = P_\zeta(\zeta_1) = P_X(x_1) \wedge \alpha + P_X(x_2) \wedge \alpha$$
  $$> \alpha, \quad \text{if } P_X(x_1) > \alpha \text{ or } P_X(x_2) > \alpha.$$

  Thus, for any Q_X such that $\{P_X \in \mathcal{P}(\mathcal{V}) : \mathsf{D_{TV}}(P_X, Q_X) \leq \epsilon\} \subseteq \{P_X \in \mathcal{P}(\mathcal{V}) : P_X(x_1) > \alpha \text{ or } P_X(x_2) > \alpha\}$, the false-alarm constraint is violated when (12) holds with equality.

If we consider a detector γ as follows: $\gamma(x_1,\zeta_1)=\gamma(x_2,\zeta_1)=\gamma(x_2,\zeta_2)=1$ and $\gamma(x,\zeta)=0$ for all other pairs $(x,\zeta)\in\mathcal{V}\times\mathcal{Z}$. We still have $\Xi(x_1)\cap\Xi(x_2)=\{\zeta_1\}$. When (12) holds with equality, i.e.,

$$P_{X,\zeta}(x_1,\zeta_1)=P_X(x_1)\wedge\alpha \quad \text{and} \quad P_{X,\zeta}(x_2,\zeta_1)+P_{X,\zeta}(x_2,\zeta_2)=P_X(x_2)\wedge\alpha,$$

we have the worst-case Type-I error lower bounded by

$$\sup_{Q_X}\left(Q_X(x_1)P_\zeta(\zeta_1)+Q_X(x_2)(P_\zeta(\zeta_1)+P_\zeta(\zeta_2))\right)=\sup_{Q_X}\left(P_\zeta(\zeta_1)+Q_X(x_2)P_\zeta(\zeta_2)\right)$$
$$=P_\zeta(\zeta_1)+P_\zeta(\zeta_2)=P_X(x_1)\wedge\alpha+P_X(x_2)\wedge\alpha>\alpha, \quad \text{if } P_X(x_1)>\alpha \text{ or } P_X(x_2)>\alpha,$$

which is the same as the previous result.

If we let $\mathcal{V}=\{x_1,x_2,x_3\}$, $\mathcal{Z}=\{\zeta_1,\zeta_2,\zeta_3,\zeta_4\}$ and $\gamma(x_3,\zeta_3)=1$ in addition to the aforementioned γ, we can similarly show that the worst-case Type-I error is larger than α for some distributions Q_X.

Therefore, it can be observed that as long as $\Xi(x_1^T)\cap\Xi(y_1^T)\neq\emptyset$ for some $x_1^T\neq y_1^T\in\mathcal{V}^T$, (12) can not be achieved with equality for all $Q_{X_1^T}$ and ϵ even for larger $(T,\mathcal{V},\mathcal{Z})$ as well as continuous $\mathcal{Z}$.

In conclusion, for any detector $\gamma\notin\Gamma^*$, the universal minimum Type-II error in (2) cannot be obtained for all $Q_{X_1^T}$ and ϵ.

Since the optimal detector takes the form $\gamma(X_1^T,\zeta_1^T)=\mathbb{1}\{X_1^T=g(\zeta_1^T)\}$ for some surjective function $g:\mathcal{Z}^T\to\mathcal{S}$, $\mathcal{S}\supset\mathcal{V}^T$, and the token vocabulary is discrete, it suffices to consider discrete $\mathcal{Z}$ to derive the optimal watermarking scheme.

Under the watermarking scheme $P^*_{X_1^T,\zeta_1^T}$ (cf. (10) and (13)), the Type-I and Type-II errors are given by:

Type-I error:

$$\forall y_1^T\in\mathcal{V}^T,\quad \mathbb{E}_{P^*_{\zeta_1^T}}[\mathbb{1}\{y_1^T=g(\zeta_1^T)\}]=\sum_{\zeta_1^T}P^*_{\zeta_1^T}(\zeta_1^T)\mathbb{1}\{y_1^T=g(\zeta_1^T)\}$$
$$=\sum_{\zeta_1^T}\sum_{x_1^T}P^*_{X_1^T,\zeta_1^T}(x_1^T,\zeta_1^T)\mathbb{1}\{y_1^T=g(\zeta_1^T)\}$$
$$=P^*_{X_1^T}(y_1^T)\sum_{\zeta_1^T}P^*_{\zeta_1^T|X_1^T}(\zeta_1^T|y_1^T)\mathbb{1}\{y_1^T=g(\zeta_1^T)\}=P^*_{X_1^T}(y_1^T)\wedge\alpha$$
$$\leq\alpha,$$

and since any distribution $Q_{X_1^T}$ can be written as a linear combinations of $\delta_{y_1^T}$, we have

$$\max_{Q_{X_1^T}}\mathbb{E}_{Q_{X_1^T}P^*_{\zeta_1^T}}[\mathbb{1}\{X_1^T=g(\zeta_1^T)\}]\leq\alpha.$$

Type-II error:

$$1-\mathbb{E}_{P^*_{X_1^T,\zeta_1^T}}[\mathbb{1}\{X_1^T=g(\zeta_1^T)\}]$$
$$=1-\sum_{x_1^T}\sum_{\zeta_1^T}P^*_{X_1^T,\zeta_1^T}(x_1^T,\zeta_1^T)\mathbb{1}\{x_1^T=g(\zeta_1^T)\}$$
$$=1-\sum_{x_1^T}P^*_{X_1^T}(x_1^T)\sum_{\zeta_1^T}P^*_{\zeta_1^T|X_1^T}(\zeta_1^T|x_1^T)\mathbb{1}\{x_1^T=g(\zeta_1^T)\}$$
$$=1-\sum_{x_1^T}\left(P^*_{X_1^T}(x_1^T)\wedge\alpha\right)$$
$$=\sum_{x_1^T:P^*_{X_1^T}(x_1^T)>\alpha}(P^*_{X_1^T}(x_1^T)-\alpha).$$

The optimality of $P^*_{X_1^T,\zeta_1^T}$ is thus proved. We note that (12) in (Opt-II) holds with equality under this optimal conditional distribution $P^*_{\zeta_1^T|X_1^T}$.

Compared to Huang et al. (2023, Theorem 3.2), their proposed detector is equivalent to $\gamma(X_1^T,\zeta_1^T) = \mathbb{1}\{X_1^T = \zeta_1^T\}$, where $\mathcal{Z}^T = \mathcal{V}^T \cup \{\tilde{\zeta}_1^T\}$ and $\tilde{\zeta}_1^T \notin \mathcal{V}^T$, meaning that it belongs to Γ^*. □

F Illustration of Construction of the Optimal Watermarking Scheme

Using a toy example in Figure 5, we now illustrate how to construct the optimal watermarking schemes, where

$$P^*_{X_1^T} = \underset{P_{X_1^T}:\mathsf{D}(P_{X_1^T},Q_{X_1^T})\le\epsilon}{\arg\min} \sum_{x_1^T}(P_{X_1^T}(x_1^T)-\alpha)_+.$$

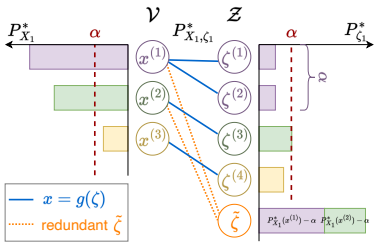

Figure 5: A toy example of the optimal detector and watermarking scheme. Links between $\mathcal{V}$ and $\mathcal{Z}$ suggest $P^*_{X_1,\zeta_1} > 0$.

Constructing the optimal watermarking scheme $P^*_{X_1^T,\zeta_1^T}$ is equivalent to transporting the probability mass $P^*_{X_1^T}$ on $\mathcal{V}$ to $\mathcal{Z}$, maximizing $P^*_{X_1^T,\zeta_1^T}(x_1^T,\zeta_1^T)$ when $x_1^T = g(\zeta_1^T)$, while keeping the worst-case Type-I error below α. Without loss of generality, by letting $T=1$, we present Figure 5 to visualize the optimal watermarking scheme. The construction process is given step by step as follows:

– **Identify text-auxiliary pairs:** We begin by identifying text-auxiliary pairs $(x,\zeta) \in \mathcal{V}\times\mathcal{Z}$ with $\gamma(x,\zeta) = \mathbb{1}\{x = g(\zeta)\} = 1$ and connect them by blue solid lines.

– **Introducing redundant auxiliary value:** We enlarge $\mathcal{Z}$ to include an additional value $\tilde{\zeta}$ and set $\gamma(x,\tilde{\zeta}) = 0$ for all x. We will call $\tilde{\zeta}$ "redundant".

– **Mass allocation for $P^*_{X_1}(x) > \alpha$:** If $P^*_{X_1}(x) > \alpha$, we transfer α mass of $P^*_{X_1}(x)$ to the ζ connected by the blue solid lines. The excess mass is transferred to the redundant $\tilde{\zeta}$ (orange dashed lines). Specifically, for $x^{(1)}$, where $P^*_{X_1}(x^{(1)}) > \alpha$ and $x^{(1)} = g(\zeta^{(1)}) = g(\zeta^{(2)})$, we move α units of mass from $P^*_{X_1}(x^{(1)})$ to $P^*_{\zeta_1}(\zeta^{(1)})$ and $P^*_{\zeta_1}(\zeta^{(2)})$, ensuring that $P^*_{\zeta_1}(\zeta^{(1)}) + P^*_{\zeta_1}(\zeta^{(2)}) = \alpha$. The rest $(P^*_{X_1}(x^{(1)}) - \alpha)$ units of mass is moved to $\tilde{\zeta}$. Similarly, for $x^{(2)}$, where $P^*_{X_1}(x^{(2)}) > \alpha$ and $x^{(2)} = g(\zeta^{(3)})$, we move α mass from $P^*_{X_1}(x^{(2)})$ to $P^*_{\zeta_1}(\zeta^{(3)})$ and $(P^*_{X_1}(x^{(2)}) - \alpha)$ mass to $\tilde{\zeta}$. Consequently, the probability of $\tilde{\zeta}$ is $P_{\zeta_1}(\tilde{\zeta}) = (P^*_{X_1}(x^{(1)}) - \alpha) + (P^*_{X_1}(x^{(2)}) - \alpha)$. In this way, there is a chance for the lower-entropy texts $x^{(1)}$ and $x^{(2)}$ to be mapped to the redundant $\tilde{\zeta}$ during watermark generation.

– **Mass allocation for $P^*_{X_1}(x) < \alpha$:** For $x^{(3)}$, where $P^*_{X_1}(x^{(3)}) < \alpha$ and $x^{(3)} = g(\zeta^{(4)})$, we move the entire mass $P^*_{X_1}(x^{(3)})$ to $P^*_{\zeta_1}(\zeta^{(4)})$ along the blue solid line. It means that higher-entropy texts will not be mapped to the redundant $\tilde{\zeta}$ during watermark generation.

– **Outcome:** This construction ensures that $P^*_{\zeta_1}(\zeta) \le \alpha$ for all $\zeta \in \{\zeta^{(1)},\zeta^{(2)},\zeta^{(3)},\zeta^{(4)}\}$, keeping the worst-case Type-I error under control. The Type-II error is equal to $P^*_{\zeta_1}(\tilde{\zeta})$, which is exactly the universally minimum Type-II error. This scheme can be similarly generalized to $T > 1$.

In Figure 5, when there is no link between $(x,\zeta) \in \mathcal{V}\times\mathcal{Z}$, the joint probability $P^*_{X_1,\zeta_1}(x,\zeta) = 0$. By letting $\epsilon = 0$, the scheme guarantees that the watermarked LLM remains unbiased (distortion-free).

Note that the detector proposed in Huang et al. (2023, Theorem 3.2) is also included in our framework, see Appendix E.

G Construction of Token-level Optimal Watermarking Scheme

The toke-level optimal watermarking scheme is the optimal solution to the following optimization problem:

$$\inf_{P_{X_t,\zeta_t|X_1^{t-1},\zeta_1^{t-1}}} \mathbb{E}_{P_{X_t,\zeta_t|X_1^{t-1},\zeta_1^{t-1}}}[1-\mathbb{1}\{X_t=g_{\mathrm{tk}}(\zeta_t)\}]$$
$$\text{s.t.} \sup_{Q_{X_t|X_1^{t-1}}} \mathbb{E}_{Q_{X_t|X_1^{t-1}}\otimes P_{\zeta_t|\zeta_1^{t-1}}}[\mathbb{1}\{X_t=g_{\mathrm{tk}}(\zeta_t)\}]\le\eta,\ \mathsf{D}_{\mathsf{TV}}(P_{X_t|X_1^{t-1}},Q_{X_t|X_1^{t-1}})\le\epsilon.$$

The optimal solution $P^*_{X_t,\zeta_t|X_1^{t-1},\zeta_1^{t-1}}$ follows the similar rule as that of $P^*_{X_1^T,\zeta_1^T}$ in Theorem 2 with $(Q_{X_1^T},P_{X_1^T},\alpha)$ replaced by $(Q_{X_t|X_1^{t-1}},P_{X_t|X_1^{t-1}},\eta)$. We refer readers to Appendix E for further details.

H Formal Statement of Lemma 3 and its Proof

Let $P^{\mathrm{token}*}_{X_1^T,\zeta_1^T}$ and $P^{\mathrm{token}*}_{\zeta_1^T}$ denote the joint distributions induced by the token-level optimal watermarking scheme.

Lemma 3 (Formal) (Token-level optimal watermarking detection errors)**.** *Let $\eta=(\alpha/\binom{T}{\lceil T\lambda\rceil})^{\frac{1}{\lceil T\lambda\rceil}}$. Under the detector γ in* (5) *and the token-level optimal watermarking scheme $P^*_{X_t,\zeta_t|X_1^{t-1},\zeta_1^{t-1}}$, the Type-I error is upper bounded by*

$$\sup_{Q_{X_1^T}}\beta_0(\gamma,Q_{X_1^T},P^{\mathrm{token}*}_{\zeta_1^T})\le\alpha.$$

Assume that when T and $n\le T$ are both large enough, token X_t is independent of X_{t-i}, i.e., $P_{X_t,X_{t-i}}=P_{X_t}\otimes P_{X_{t-i}}$, for all $i\ge n+1$ and $t\in[T]$. Let $\mathcal{I}_{T,n}(i)=([i-n,i+n]\cap[T])\backslash\{i\}$. By setting the detector threshold as $\lambda=\frac{a}{T}\sum_{t=1}^T\mathbb{E}_{X_t,\zeta_t}[\mathbb{1}\{X_t=g(\zeta_t)\}]$ for some $a\in[0,1]$, the Type-II error exponent is

$$-\log\beta_1(\gamma,P^{\mathrm{token}*}_{X_1^T,\zeta_1^T})=\Omega\left(\frac{T}{n}\right).$$

The following is the proof of Lemma 3.

To choose $\lceil T\lambda\rceil$ indices out of $\{1,\dots,T\}$, there are $\binom{T}{\lceil T\lambda\rceil}$ choices. Let $k=1,\dots,\binom{T}{\lceil T\lambda\rceil}$ and S_k be the k-th set of the chosen indices. The Type-I error is upper bounded by

$$\begin{aligned}\beta_0(\gamma,Q_{X^{(T)}},P^{\mathrm{token}*}_{\zeta_1^T})&=\Pr\left(\frac{1}{T}\sum_{t=1}^T\mathbb{1}\{X_t=g(\zeta_t)\}\ge\lambda\mid \mathrm{H}_0\right)\\&\le\Pr\left(\bigcup_{k=1}^{\binom{T}{\lceil T\lambda\rceil}}\{\mathbb{1}\{X_t=g(\zeta_t)\}=1,\forall t\in S_k\}\mid\mathrm{H}_0\right)\\&\le\sum_{k=1}^{\binom{T}{\lceil T\lambda\rceil}}\underbrace{\Pr\left(\{\mathbb{1}\{X_t=g(\zeta_t)\}=1,\forall t\in S_k\}\mid\mathrm{H}_0\right)}_{P_{\mathrm{FA},k}}.\end{aligned}$$

Without loss of generality, let $m=\lceil T\lambda\rceil$ and $S_k=\{1,2,\dots,m\}$. We can rewrite $P_{\mathrm{FA},k}$ as

$$\begin{aligned}P_{\mathrm{FA},k}&=\mathbb{E}_{Q_{X^{(T)}}\otimes P_{\zeta^{(T)}}}[\{\mathbb{1}\{X_t=g(\zeta_t)\}=1,\forall t\in S_k\}]\\&=\mathbb{E}_{Q_{X^{(T)}}\otimes P_{\zeta^{(T)}}}[\prod_{t\in S_k}\mathbb{1}\{X_t=g(\zeta_t)\}]\end{aligned}$$

$$= \mathbb{E}_{Q_{X_1}\otimes P_{\zeta_1}}\Bigg[\mathbb{1}\{X_1 = g(\zeta_1)\}\mathbb{E}_{Q_{X_2|X_1}\otimes P_{\zeta_2|\zeta_1}}\Bigg[\mathbb{1}\{X_2 = g(\zeta_2)\}\cdots \cdots \mathbb{E}_{Q_{X_m|X_1^{m-1}}\otimes P_{\zeta_m|\zeta_1^{m-1}}}[\mathbb{1}\{X_m = g(\zeta_m)\}]\big]\cdots\Bigg]\Bigg]$$

$$\leq \eta^m, \quad \forall Q_{X_1^T}.$$

Then the Type-I error is finally upper bounded by

$$\sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}^{\text{token}*}) \leq \binom{T}{\lceil T\lambda\rceil}\eta^{\lceil T\lambda\rceil} \leq \alpha.$$

We prove the Type-II error bound by applying Janson (1998, Theorem 10).

Theorem 5 (Theorem 10, Janson (1998))**.** *Let $\{I_i\}_{i\in\mathcal{I}}$ be a finite family of indicator random variables, defined on a common probability space. Let G be a dependency graph of $\mathcal{I}$, i.e., a graph with vertex set $\mathcal{I}$ such that if A and B are disjoint subsets of $\mathcal{I}$, and Γ contains no edge between A and B, then $\{I_i\}_{i\in A}$ and $\{I_i\}_{i\in B}$ are independent. We write $i \sim j$ if $i, j \in \mathcal{I}$ and (i, j) is an edge in G. In particular, $i \not\sim i$. Let $S = \sum_{i\in\mathcal{I}} I_i$ and $\Delta = \mathbb{E}[S]$. Let $\Psi = \max_{i\in\mathcal{I}}\sum_{j\in\mathcal{I},j\sim i}\mathbb{E}[I_j]$ and $\Phi = \frac{1}{2}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I},j\sim i}\mathbb{E}[I_iI_j]$. For any $0 \leq a \leq 1$,*

$$\Pr(S \leq a\Delta) \leq \exp\left\{-\min\left\{(1-a)^2\frac{\Delta^2}{8\Phi + 2\Delta}, (1-a)\frac{\Delta}{6\Psi}\right\}\right\}. \tag{14}$$

Given any detector γ that accepts the form in (5) and the corresponding optimal watermarking scheme, for some $a \in (0, 1)$, we first set the threshold in γ as

$$T\lambda = a\sum_{t=1}^{T}\mathbb{E}_{X_t,\zeta_t}[\mathbb{1}\{X_t = g(\zeta_t)\}] = a\sum_{t=1}^{T}\mathbb{E}_{X_1^{t-1}}\left[\sum_x \left(P^*_{X_t|X_1^{t-1}}(x|X_1^{t-1}) - \eta\right)_+\right] =: a\Delta_T,$$

where $P^*_{X_t|X_1^{t-1}}$ is induced by $P^*_{X_t,\zeta_t|X_1^{t-1},\zeta_1^{t-1}}$. The Type-II error is given by

$$\beta_1(\gamma, P^{\text{token}*}_{X_1^T,\zeta_1^T}) = P^{\text{token}*}_{X_1^T,\zeta_1^T}\left(\sum_{t=1}^{T}\mathbb{1}\{X_t = g(\zeta_t)\} < a\Delta_T\right)$$

which is exactly the left-hand side of (14).

Assume that when T and $n \leq T$ are large enough, token X_t is independent of all X_{t-i} for all $i \geq n+1$ and $t \in [T]$, i.e., $P_{X_t,X_{t-i}} = P_{X_t}\otimes P_{X_{t-i}}$. Let $\mathcal{I}_{T,n}(i) = ([i-n, i+n]\cap[T])\backslash\{i\}$. The Ψ and Φ on the right-hand side of (14) are given by:

$$\Psi \coloneqq \max_{i\in[T]}\sum_{t\in[T],t\sim i}\mathbb{E}_{X_t,\zeta_t}[\mathbb{1}\{X_t = g(\zeta_t)\}] = \max_{i\in[T]}\sum_{t\in\mathcal{I}_{T,n}(i)}\mathbb{E}_{X_t,\zeta_t}[\mathbb{1}\{X_t = g(\zeta_t)\}] = \Theta(n),$$

$$\Phi \coloneqq \frac{1}{2}\sum_{i\in[T]}\sum_{j\in[T],j\sim i}\mathbb{E}[\mathbb{1}\{X_i = g(\zeta_i)\}\mathbb{1}\{X_j = g(\zeta_j)\}]$$

$$= \frac{1}{2}\sum_{i\in[T]}\sum_{j\in\mathcal{I}_{T,n}(i)}\mathbb{E}[\mathbb{1}\{X_i = g(\zeta_i)\}\mathbb{1}\{X_j = g(\zeta_j)\}] = \Theta(Tn).$$

By plugging Δ_T, Ω and Θ back into the right-hand side of (14), we have the upper bound

$$\beta_1(\gamma, P^{\text{token}*}_{X_1^T,\zeta_1^T}) \leq \exp\left\{-\min\left\{(1-a)^2\frac{\Delta_T^2}{8\Phi + 2\Delta_T}, (1-a)\frac{\Delta_T}{6\Psi}\right\}\right\}$$

where $U_t = \mathbb{E}_{X_1^{t-1}}\big[\sum_x \left(P^*_{X_t|X_1^{t-1}}(x|X_1^{t-1}) - \eta\right)_+\big]$, $\Delta_T \coloneqq \sum_{t=1}^{T}U_t$, $\Psi = \max_{i\in[T]}\sum_{t\in\mathcal{I}_{T,n}(i)}U_t$, and $\Phi = \frac{1}{2}\sum_{i\in[T]}\sum_{j\in\mathcal{I}_{T,n}(i)}\mathbb{E}[\mathbb{1}\{X_i = g(\zeta_i)\}\mathbb{1}\{X_j = g(\zeta_j)\}]$. This implies

$$-\log\beta_1(\gamma, P^{\text{token}*}_{X_1^T,\zeta_1^T}) \geq \min\left\{(1-a)^2\Theta\left(\frac{T}{n}\right), (1-a)\Theta\left(\frac{T}{n}\right)\right\}$$

$$\Longrightarrow -\log\beta_1(\gamma, P^{\text{token}*}_{X_1^T,\zeta_1^T}) = \Omega\left(\frac{T}{n}\right).$$

I Optimal Watermarking Scheme with Uniform $P_{\zeta_1^T}$ for $\gamma \in \Gamma^*$

With any detector $\gamma \in \Gamma^*$ from the class of optimal detectors proposed in Theorem 2, we consider constructing an optimal watermarking scheme with fixed the marginal distribution $P_{\zeta_1^T} = \mathrm{Unif}(\mathcal{Z}^T)$. This simplifies the transmission of ζ_1^T to the detector by using a shared key to sample ζ_1^T via a hash function.

The optimization problem is (Opt-I), where we choose the distortion metric as $\mathsf{D_{TV}}$ and aim to solve the joint distribution $P_{X_1^T,\zeta_1^T}$ that maximizes the detection performance with distortion guarantee. The alternative watermarking scheme optimal for $\gamma \in \Gamma^*$ when g is an identity mapping is given in the following lemma. Note that this result can be generalized to other functions g.

Lemma 6 (Optimal watermarking scheme for $\gamma = \mathbb{1}\{X_1^T = \zeta_1^T\}$ when $P_{\zeta_t} = \mathrm{Unif}(\mathcal{Z})$)**.** *When $\gamma = \mathbb{1}\{X_1^T = \zeta_1^T\}$, $P_{\zeta_t} = \mathrm{Unif}(\mathcal{Z})$, and $\alpha \geq \frac{1}{|\mathcal{Z}|^T}$, the minimum Type-II error is $\min_{P_{X_1^T}:\mathsf{D_{TV}}(P_{X_1^T}\|Q_{X_1^T})\leq\epsilon} \sum_{x_1^T} \left(P_{X_1^T}(x_1^T) - \frac{1}{|\mathcal{Z}|^T}\right)_+$. The optimal ϵ-distorted watermarking scheme that achieves the minimum Type-II error is*

$$P^*_{X_1^T,\zeta_1^T}(x_1^T,\zeta_1^T) = \begin{cases} \min\{P^*_{X_1^T}(x_1^T), \frac{1}{|\mathcal{Z}|^T}\}, & \text{if } x_1^T = \zeta_1^T; \\ \frac{\left(P^*_{X_1^T}(x_1^T) - \frac{1}{|\mathcal{Z}|^T}\right)_+ \cdot \left(\frac{1}{|\mathcal{Z}|^T} - P^*_{X_1^T}(\zeta_1^T)\right)_+}{\mathsf{D_{TV}}(P^*_{X_1^T}, \mathrm{Unif}(\mathcal{Z}^T))}, & \text{otherwise}, \end{cases}$$

where $P^*_{X_1^T} = \arg\min_{P_{X_1^T}:\mathsf{D_{TV}}(P_{X_1^T}\|Q_{X_1^T})\leq\epsilon} \sum_{x_1^T} \left(P_{X_1^T}(x_1^T) - \frac{1}{|\mathcal{Z}|^T}\right)_+$.

The proof of Lemma 6 follows from the fact that $\mathsf{D_{TV}}(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \pi(X \neq Y)$ Thorisson (1995), where $X \sim \mu$, $Y \sim \nu$ and $\Pi(\mu,\nu)$ is the set of all couplings of Borel probability measures μ and ν. Note that when $P_{\zeta_t} = \mathrm{Unif}(\mathcal{Z})$, if $\alpha < \frac{1}{|\mathcal{Z}|^T}$, the feasible region of (Opt-I) becomes empty. With this watermarking scheme, the detector can fully recover ζ_1^T using a pseudorandom generator and shared key. However, the resulting minimum Type-II error is larger than $\beta_1^*(Q_{X_1^T}, \alpha, \epsilon)$ from Theorem 1, as $\alpha \geq \frac{1}{|\mathcal{Z}|^T}$. In practice, the gap is significant since $\frac{1}{|\mathcal{Z}|^T} = \mathcal{O}(10^{-4T})$ is much smaller than typical values of α. This gap reflects the cost of pseudo-transmitting ζ_1^T using only the shared key. Nonetheless, if $T = 1$, it is possible to set false alarm constraint to $\alpha = \frac{1}{|\mathcal{Z}|}$ and mitigate the performance loss. Motivated by this, we move on to discuss the token-level optimal watermarking scheme in the subsequent sections.

Proof. Consider $\gamma(X_1^T,\zeta_1^T) = \mathbb{1}\{X_1^T = \zeta_1^T\}$ and $\mathcal{V}^T \subseteq \mathcal{Z}^T$, which is a model-agnostic detector. We first start the proof from the distortion-free setting with $\epsilon = 0$ and an arbitrary distribution $P_{\zeta_1^T}$ on $\mathcal{Z}^T$. The objective function (i.e. Type-II error) becomes $P_{X_1^T,\zeta_1^T}(X_1^T \neq \zeta_1^T)$, whose minimum is well-known as $\mathsf{D_{TV}}(Q_{X_1^T}, P_{\zeta_1^T})$ and the minimizer is

$$P^*_{X_1^T,\zeta_1^T}(x_1^T,\zeta_1^T) = \begin{cases} \min\{Q_{X_1^T}(x_1^T), P_{\zeta_1^T}(\zeta_1^T)\}, & \text{if } x_1^T = \zeta_1^T; \\ \frac{(Q_{X_1^T}(x_1^T) - P_{\zeta_1^T}(x_1^T))_+ \cdot (P_{\zeta_1^T}(\zeta_1^T) - Q_{X_1^T}(\zeta_1^T))_+}{\mathsf{D_{TV}}(Q_{X_1^T}, P_{\zeta_1^T})}, & \text{otherwise.} \end{cases} \tag{15}$$

This holds for any given pair of $(Q_{X_1^T}, P_{\zeta_1^T})$. This watermarking scheme basically tries to force $X_1^T = \zeta_1^T$ as often as possible. However, we need to design $P_{\zeta_1^T}$ such that the Type-I error probability $\sup_{Q_{X_1^T}} \mathbb{E}_{Q_{X_1^T} P_{\zeta_1^T}}[\mathbb{1}\{X_1^T = \zeta_1^T\}] \leq \alpha$, i.e.,

$$\begin{aligned} P^*_{\zeta_1^T} &\coloneqq \underset{P_{\zeta_1^T}:\, \sup_{Q_{X_1^T}} \mathbb{E}_{Q_{X_1^T}\otimes P_{\zeta_1^T}}[\mathbb{1}\{X_1^T=\zeta_1^T\}]\leq\alpha}{\arg\min} \mathsf{D_{TV}}(Q_{X_1^T}, P_{\zeta_1^T}) \\ &= \underset{P_{\zeta_1^T}:\, \sup_{Q_{X_1^T}} \langle Q_{X_1^T}, P_{\zeta_1^T}\rangle\leq\alpha}{\arg\min} \sum_{x_1^T\in\mathcal{V}^T} \left(Q_{x_1^T}(x_1^T) - P_{\zeta_1^T}(x_1^T)\right)_+. \end{aligned}$$

To further consider cases where we allow distortion $\mathsf{D}(P_{X_1^T}\|Q_{X_1^T}) \le \epsilon$ for some $\epsilon \ge 0$, we solve

$$\begin{aligned}(P^*_{X_1^T}, P^*_{\zeta_1^T}) \coloneqq & \underset{\substack{(P_{X_1^T}, P_{\zeta_1^T}):\\ \mathsf{D}_{\mathsf{TV}}(P_{X_1^T}\|Q_{X_1^T})\le\epsilon,\\ \sup_{Q_{X_1^T}} \langle Q_{X_1^T}, P_{\zeta_1^T}\rangle \le \alpha}}{\arg\min} \mathsf{D}(P_{X_1^T}, P_{\zeta_1^T}) \\ = & \underset{\substack{(P_{X_1^T}, P_{\zeta_1^T}):\\ \mathsf{D}(P_{X_1^T}\|Q_{X_1^T})\le\epsilon,\\ \sup_{Q_{X_1^T}} \langle Q_{X_1^T}, P_{\zeta_1^T}\rangle \le \alpha}}{\arg\min} \sum_{x_1^T \in \mathcal{V}^T} \left(P_{x_1^T}(x_1^T) - P_{\zeta_1^T}(x_1^T)\right)_+,\end{aligned}$$

and plug them into (15). We then move to the special case when $P_{\zeta_1^T}$ is uniform.

Special case ($\mathcal{V}^T \subseteq \mathcal{S} \subseteq \mathcal{Z}^T$ and $P_{\zeta_1^T} = \mathrm{Unif}(\mathcal{S})$). For any $\zeta_1^T \in \mathcal{S}$, $P_{\zeta_1^T}(\zeta_1^T) = \frac{1}{|\mathcal{S}|}$. To ensure that the false alarm constraint is satisfied, we require $\alpha \ge \sup_{Q_{X_1^T}} \sum_{x_1^T} Q_{X_1^T}(x_1^T) \cdot \frac{1}{|\mathcal{S}|} = \frac{1}{|\mathcal{S}|}$. In other words, to enforce lower false alarm probability, we need to increase the size of $\mathcal{S}$. The minimum Type-II error probability is given by

$$\mathsf{D}_{\mathsf{TV}}(Q_{X_1^T}, \mathrm{Unif}(\mathcal{S})) = \sum_{x_1^T \in \mathcal{V}^T} \left(Q_{X_1^T}(x_1^T) - \frac{1}{|\mathcal{S}|}\right)_+.$$

If $|\mathcal{S}| = \frac{1}{\alpha}$, this minimum Type-II error is equal to the optimal result $\sum_{x_1^T \in \mathcal{V}^T}(Q_{X_1^T}(x_1^T) - \alpha)_+$. Otherwise, if $|\mathcal{S}| > \frac{1}{\alpha}$, this Type-II error is larger and the gap represents the price paid by using the uniform distribution $P_{\zeta_1^T}$, i.e., sending pseudorandom numbers.

□

J DAWA PSEUDO-CODES

Algorithm 1 Watermarked Text Generation

Input: LLM Q, Vocabulary $\mathcal{V}$, Prompt u, Secret $\mathtt{key}$, Token-level false alarm η.

$\mathcal{Z} = \{h_{\mathtt{key}}(x)\}_{x\in\mathcal{V}} \cup \{\tilde{\zeta}\}$
for $t = 1, ..., T$ **do**
  $P_{\zeta_t|x_1^{t-1},u}(\zeta) \leftarrow (Q_{X_t|x_1^{t-1},u}(h^{-1}_{\mathtt{key}}(\zeta)) \wedge \eta),\ \forall \zeta \in \mathcal{Z}\backslash\{\tilde{\zeta}\}$.
  $P_{\zeta_t|x_1^{t-1},u}(\tilde{\zeta}) \leftarrow \sum_{x\in\mathcal{V}}(Q_{X_t|x_1^{t-1},u}(x) - \eta)_+$.
  Compute a hash of tokens x_{t-n}^{t-1} with $\mathtt{key}$, and use it as a seed to generate $(G_{t,\zeta})_{\zeta\in\mathcal{Z}}$ from Gumbel distribution.
  $\zeta_t \leftarrow \arg\max_{\zeta\in\mathcal{Z}} \log(P_{\zeta_t|x_1^{t-1},u}(\zeta)) + G_{t,\zeta}$.
  if $\zeta_t \neq \tilde{\zeta}$ **then**
    $x_t \leftarrow h^{-1}_{\mathtt{key}}(\zeta_t)$
  else
    Sample $x_t \sim \left(\frac{(Q_{X_t|x_1^{t-1},u}(x)-\eta)_+}{\sum_{x\in\mathcal{V}}\left(Q_{X_t|x_1^{t-1},u}(x)-\eta\right)_+}\right)_{x\in\mathcal{V}}$
  end if
end for

Output: Watermarked text $x_1^T = (x_1, ..., x_T)$.

Algorithm 2 Watermarked Text Detection

Input: SLM $\tilde{Q}$, Vocabulary $\mathcal{V}$, Text x_1^T, Secret $\texttt{key}$, Token-level false alarm η, Threshold λ.

score $= 0$, $\mathcal{Z} = \{h_{\texttt{key}}(x)\}_{x\in\mathcal{V}} \cup \{\tilde{\zeta}\}$
for $t = 1, ..., T$ **do**
  $\tilde{P}_{\zeta_t|x_1^{t-1}}(\zeta) \leftarrow (\tilde{Q}_{X_t|x_1^{t-1}}(h_{\texttt{key}}^{-1}(\zeta)) \wedge \eta),\ \forall \zeta \in \mathcal{Z}\backslash\{\tilde{\zeta}\}$.
  $\tilde{P}_{\zeta_t|x_1^{t-1}}(\tilde{\zeta}) \leftarrow \sum_{x\in\mathcal{V}} (\tilde{Q}_{X_t|x_1^{t-1}}(x) - \eta)_+$.
  Compute a hash of tokens x_{t-n}^{t-1} with $\texttt{key}$, and use it as a seed to generate $(G_{t,\zeta})_{\zeta\in\mathcal{Z}}$ from Gumbel distribution.
  $\zeta_t \leftarrow \arg\max_{\zeta\in\mathcal{Z}} \log(\tilde{P}_{\zeta_t|x_1^{t-1}}(\zeta)) + G_{t,\zeta}$.
  score $\leftarrow$ score $+ \mathbb{1}\{h_{\texttt{key}}(x_t) = \zeta_t\}$
end for
if score $> T\lambda$ **then**
  return 1 ▷ Input text is watermarked
else
  return 0 ▷ Input text is unwatermarked
end if

K Additional Experimental Results

Empirical analysis on False Alarm Control. We conduct experiments to show the relationship between theoretical FPR (i.e., α) and the corresponding empirical FPR. As discussed in Lemma 3, we set the token-level false alarm rate as $\eta = 0.1$ and the sequence length as $T = 200$, which controls the sequence-level false alarm rate under $\alpha = \binom{T}{\lceil T\lambda \rceil}\eta^{\lceil T\lambda \rceil}$, where λ is the detection threshold. For a given theoretical FPR α, we calculate the corresponding threshold λ and the empirical FPR based on 100k unwatermarked sentences. The results, as shown in Table 4, confirm that our theoretical guarantee effectively controls the empirical false alarm rate.

Table 4: Theoretical and empirical FPR under different thresholds.

Theoretical FPR	9e-03	2e-03	5e-04	9e-05
Empirical FPR	1e-04	4e-05	2e-05	2e-05

Efficiency of Watermark Scheme. To evaluate the efficiency of our watermarking method, we conduct experiments to measure the average generation time for both watermarked and unwatermarked text. In both scenarios, we generated 500 texts, each containing 200 tokens. Table 5 indicates that the difference in generation time between unwatermarked and watermarked text is less than 0.5 seconds. This minimal difference confirms that our watermarking method has a negligible impact on generation speed, ensuring practical applicability.

Table 5: Average generation time comparison for watermarked and unwatermarked text using Llama2-13B.

Language Model	**Setting**	**Avg Generation Time (s)**
Llama2-13B	Unwatermarked	9.110
Llama2-13B	Watermarked	9.386

Surrogate Language Model. SLM plays a crucial role during the detection process of our watermarking method. We examine how the choice of SLM affects the detection performance of our watermarking scheme. The selection of the surrogate model is primarily based on its vocabulary or tokenizer rather than the specific language model within the same family. This choice is critical because, during detection, the text must be tokenized exactly using the same tokenizer as the watermarking model to ensure accurate token recovery. As a result, any language model that employs the same tokenizer can function effectively as the surrogate model. To validate our approach, we apply

our watermarking algorithm to GPT-J-6B (a model with 6 billion parameters) and use GPT-2 Large (774 million parameters) as the SLM. Despite differences in developers, training data, architecture, and training methods, these two models share the same tokenizer, making them compatible for this task. We conduct experiments using the C4 dataset, and the results are presented in Table 6. The results demonstrate the effectiveness of our proposed watermarking method with or without attack even when using a surrogate model from a different family than the watermarking language model. Notably, the surrogate model, despite having fewer parameters and lower overall capability compared to the watermarking language model, does not compromise the watermarking performance.

Table 6: Performance comparison of different language models and surrogate models under two scenarios: without attack and with token replacement attack.

Scenario	Language Model	Surrogate Model	ROC-AUC	TPR@1% FPR	TPR@10% FPR
Without Attack	Llama2-13B	Llama2-7B	0.999	0.998	1.000
	Mistral-8 × 7B	Mistral-7B	0.999	0.998	1.000
	GPT-J-6B	GPT-2 large	0.997	0.990	0.997
With Attack	Llama2-13B	Llama2-7B	0.989	0.860	0.976
	Mistral-8 × 7B	Mistral-7B	0.990	0.881	0.966
	GPT-J-6B	GPT-2 large	0.987	0.892	0.962

Prompt Agnostic. Prompt agnosticism is a crucial property of LLM watermark detection. We investigate the impact of prompts on our watermark detection performance by conducting experiments to compare detection accuracy with and without prompts attached to the watermarked text during the detection process. The results are presented in Table 7. Notably, even when prompts are absent and the SLM cannot perfectly reconstruct the same distribution of ζ_t as in the generation process, our detection performance remains almost unaffected. This demonstrates the robustness of our watermarking method, regardless of whether a prompt is included during the detection phase.

Table 7: Performance comparison of Llama2-13B under two scenarios: without attack and with token replacement attack, with and without prompts.

Scenario	Language Model	Surrogate Model	Setting	ROC-AUC	TPR@1% FPR	TPR@10% FPR
Without Attack	Llama2-13B	Llama2-7B	Without Prompt	0.997	0.983	0.995
	Llama2-13B	Llama2-7B	With Prompt	0.998	0.989	0.996
With Attack	Llama2-13B	Llama2-7B	Without Prompt	0.977	0.818	0.953
	Llama2-13B	Llama2-7B	With Prompt	0.979	0.816	0.960

L Extension to Jointly Optimal Robust Watermarking Scheme and Detector

Thus far, we have theoretically examined the optimal detector and watermarking scheme without considering adversarial scenarios. In practice, users may attempt to modify LLM output to remove watermarks through techniques like replacement, deletion, insertion, paraphrasing, or translation. We now show that our framework can be extended to incorporating robustness against these attacks.

We consider a broad class of attacks, where the text can be altered in arbitrary ways as long as certain latent pattern, such as its *semantics*, is preserved. Specifically, let $f : \mathcal{V}^T \to [K]$ be a function that maps a sequence of tokens X_1^T to a finite latent space $[K] \subset \mathbb{N}_+$; for example, $[K]$ may index K distinct semantics clusters and f is a function extracting the semantics. Clearly, f induces an equivalence relation, say, denoted by $\equiv_f$, on $\mathcal{V}^T$, where $x_1^T \equiv_f {x'}_1^T$ if and only if $f(x_1^T) = f({x'}_1^T)$. Let $\mathcal{B}_f(x_1^T)$ be an equivalence class containing x_1^T. Under the assumption that the adversary is arbitrarily powerful except that it is unable to move any x_1^T outside its equivalent class $\mathcal{B}_f(x_1^T)$ (e.g., unable to alter the semantics of x_1^T), the "f-robust" Type-I and Type-II errors are then defined as

$$\beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}, f) \coloneqq \mathbb{E}_{Q_{X_1^T} \otimes P_{\zeta_1^T}} \left[\sup_{\tilde{x}_1^T \in \mathcal{B}_f(X_1^T)} \mathbb{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 1\} \right],$$
$$\beta_1(\gamma, P_{X_1^T, \zeta_1^T}, f) \coloneqq \mathbb{E}_{P_{X_1^T, \zeta_1^T}} \left[\sup_{\tilde{x}_1^T \in \mathcal{B}_f(X_1^T)} \mathbb{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 0\} \right].$$

Designing universally optimal f-robust detector and watermarking scheme can then be formulated as jointly minimizing the f-robust Type-II error while constraining the worst-case f-robust Type-I error, namely, solving the optimization problem

$$\inf_{\gamma, P_{X_1^T, \zeta_1^T}} \beta_1(\gamma, P_{X_1^T, \zeta_1^T}, f) \quad \text{s.t.} \ \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}, f) \leq \alpha, \ \mathsf{D}_{\mathsf{TV}}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon. \quad \text{(Opt-R)}$$

We prove the following theorem.

Theorem 7 (Universally minimum f-robust Type-II error)**.** *The universally minimum f-robust Type-II error attained from* (Opt-R) *is*

$$\beta_1^*(Q_{X_1^T}, \alpha, \epsilon, f) \coloneqq \min_{P_{X_1^T}: \mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon} \sum_{k \in [K]} \left(\left(\sum_{x_1^T : f(x_1^T) = k} P_{X_1^T}(x_1^T) \right) - \alpha \right)_+ .$$

Notably, $\beta_1^*(Q_{X_1^T}, \alpha, \epsilon, f)$ aligns with the minimum Type-II error in Example 3, which is suboptimal without an adversary but becomes optimal under the adversarial settting of (Opt-R). The gap between $\beta_1^*(Q_{X_1^T}, \alpha, \epsilon, f)$ in Theorem 7 and $\beta_1^*(Q_{X_1^T}, \alpha, \epsilon)$ in Theorem 1 reflects the cost of ensuring robustness, widening as K decreases (i.e., as perturbation strength increases), see Figure 6 in appendix for an illustration of the optimal f-robust minimum Type-II error when f is a semantic mapping. Similar to Theorem 2, we derive the optimal detector and watermarking scheme achieving $\beta_1^*(Q_{X_1^T}, \alpha, \epsilon, f)$, detailed in Appendix N. These solutions closely resemble those in Theorem 2. For implementation, if the latent space $[K]$ is significantly smaller than $\mathcal{V}^T$, applying the optimal f-robust detector and watermarking scheme becomes more effective than those presented in Theorem 2. Additionally, a similar algorithmic strategy to the one discussed in Sections 4 and 5 can be employed to address the practical challenges discussed earlier. These extensions and efficient implementations of the function f in practice are promising directions of future research.

M Proof of Theorem 7

According to the Type-I error constraint, we have $\forall x_1^T \in \mathcal{V}^T$,

$$\begin{aligned}
\alpha &\geq \max_{Q_{X_1^T}} \mathbb{E}_{Q_{X_1^T} \otimes P_{\zeta_1^T}} \left[\sup_{\tilde{x}_1^T \in \mathcal{B}_f(X_1^T)} \mathbb{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 1\} \right] \\
&\geq \mathbb{E}_{\delta_{x_1^T} \otimes P_{\zeta_1^T}} \left[\sup_{\tilde{x}_1^T \in \mathcal{B}_f(X_1^T)} \mathbb{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 1\} \right] = \mathbb{E}_{P_{\zeta_1^T}} \left[\sup_{\tilde{x}_1^T \in \mathcal{B}_f(x_1^T)} \gamma(\tilde{x}_1^T, \zeta_1^T) \right] \\
&= \sum_{\zeta_1^T} P_{\zeta_1^T}(\zeta_1^T) \sup_{\tilde{x}_1^T \in \mathcal{B}_f(x_1^T)} \gamma(\tilde{x}_1^T, \zeta_1^T).
\end{aligned}$$

For brevity, let $\mathcal{B}(k) \coloneqq \mathcal{B}_f(x_1^T)$ if $f(x_1^T) = k$. The f-robust Type-II error is equal to $1 - \mathbb{E}_{P_{X_1^T, \zeta_1^T}}[\inf_{\tilde{x}_1^T \in \mathcal{B}_f(X_1^T)} \gamma(\tilde{x}_1^T, \zeta_1^T)]$. We have

$$\begin{aligned}
&\mathbb{E}_{P_{X_1^T, \zeta_1^T}} \left[\inf_{\tilde{x}_1^T \in \mathcal{B}_f(X_1^T)} \gamma(\tilde{x}_1^T, \zeta_1^T) \right] \leq \mathbb{E}_{P_{X_1^T, \zeta_1^T}} \left[\sup_{\tilde{x}_1^T \in \mathcal{B}_f(X_1^T)} \gamma(\tilde{x}_1^T, \zeta_1^T) \right] \\
&= \sum_{k \in [K]} \underbrace{\sum_{x_1^T : f(x_1^T) = k} \sum_{\zeta_1^T} P_{X_1^T, \zeta_1^T}(x_1^T, \zeta_1^T) \sup_{\tilde{x}_1^T \in \mathcal{B}_f(x_1^T)} \gamma(\tilde{x}_1^T, \zeta_1^T)}_{C(k)},
\end{aligned}$$

where according to the f-robust Type-I error constraint, for all $k \in [K]$,

$$\begin{aligned}
C(k) &\leq \sum_{x_1^T : f(x_1^T) = k} P_{X_1^T}(x_1^T), \quad \text{and} \\
C(k) &= \sum_{\zeta_1^T} P_{\zeta_1^T}(\zeta_1^T) \sum_{x_1^T : f(x_1^T) = k} P_{X_1^T | \zeta_1^T}(x_1^T | \zeta_1^T) \sup_{\tilde{x}_1^T \in \mathcal{B}(k)} \gamma(\tilde{x}_1^T, \zeta_1^T) \\
&\leq \sum_{\zeta_1^T} P_{\zeta_1^T}(\zeta_1^T) \sup_{\tilde{x}_1^T \in \mathcal{B}(k)} \gamma(\tilde{x}_1^T, \zeta_1^T) \leq \alpha.
\end{aligned}$$

Therefore,

$$\mathbb{E}_{P_{X_1^T,\zeta_1^T}}\left[\inf_{\tilde{x}_1^T\in\mathcal{B}(f(X_1^T))}\gamma(\tilde{x}_1^T,\zeta_1^T)\right] \le \sum_{k\in[K]} C(k)$$

$$\le \sum_{k\in[K]}\left(\left(\sum_{x_1^T:f(x_1^T)=k} P_{X_1^T}(x_1^T)\right)\wedge\alpha\right) = 1-\sum_{k\in[K]}\left(\left(\sum_{x_1^T:f(x_1^T)=k} P_{X_1^T}(x_1^T)\right)-\alpha\right)_+ \quad (16)$$

where (16) is maximized by taking

$$P_{X_1^T} = P^{*,f}_{X_1^T} \coloneqq \operatorname*{arg\,min}_{P_{X_1^T}:\mathsf{D}(P_{X_1^T},Q_{X_1^T})\le\epsilon}\sum_{k\in[K]}\left(\left(\sum_{x_1^T:f(x_1^T)=k} P_{X_1^T}(x_1^T)\right)-\alpha\right)_+.$$

For any $P_{X_1^T}$, the f-robust Type-II error is lower bounded by

$$\mathbb{E}_{P_{X_1^T,\zeta_1^T}}\left[\sup_{\tilde{x}_1^T\in\mathcal{B}_f(X_1^T)}\mathbb{1}\{\gamma(\tilde{x}_1^T,\zeta_1^T)=0\}\right] \ge \sum_{k\in[K]}\left(\left(\sum_{x_1^T:f(x_1^T)=k} P_{X_1^T}(x_1^T)\right)-\alpha\right)_+.$$

By plugging $P^{*,f}_{X_1^T}$ into the lower bound, we obtain the universal minimum f-robust Type-II error over all possible γ and $P_{X_1^T,\zeta_1^T}$, denoted by

$$\beta_1^*(f,Q_{X_1^T},\epsilon,\alpha) \coloneqq \min_{P_{X_1^T}:\mathsf{D}(P_{X_1^T},Q_{X_1^T})\le\epsilon}\sum_{k\in[K]}\left(\left(\sum_{x_1^T:f(x_1^T)=k} P_{X_1^T}(x_1^T)\right)-\alpha\right)_+. \quad (17)$$

N Optimal Type of f-Robust Detectors and Watermarking Schemes

Theorem 8 (Optimal type of f-robust detectors and watermarking schemes)**.** *Let Γ_f^* be a collection of detectors that accept the form*

$$\gamma(X_1^T,\zeta_1^T) = \mathbb{1}\{X_1^T = g(\zeta_1^T) \text{ or } f(X_1^T) = g(\zeta_1^T)\}$$

for some function $g:\mathcal{Z}^T\to\mathcal{S}$, $\mathcal{S}\cap([K]\cup\mathcal{V}^T)\neq\emptyset$ and $|\mathcal{S}|>K$. If an only if the detector $\gamma\in\Gamma_f^*$, the minimum Type-II error attained from (Opt-R) *reaches $\beta_1^*(Q_{X_1^T},\epsilon,\alpha,f)$ in* (17) *for all text distribution $Q_{X_1^T}\in\mathcal{P}(\mathcal{V}^T)$ and distortion level $\epsilon\in\mathbb{R}_{\ge0}$.*

After enlarging $\mathcal{Z}^T$ to include redundant auxiliary values, the ϵ-distorted optimal f-robust watermarking scheme $P^{*,f}_{X_1^T,\zeta_1^T}(x_1^T,\zeta_1^T)$ is given as follows:

$$P^{*,f}_{X_1^T} \coloneqq \operatorname*{arg\,min}_{P_{X_1^T}:\mathsf{D}_{\mathsf{TV}}(P_{X_1^T},Q_{X_1^T})\le\epsilon}\sum_{k\in[K]}\left(\left(\sum_{x_1^T:f(x_1^T)=k} P_{X_1^T}(x_1^T)\right)-\alpha\right)_+,$$

and for any $x_1^T\in\mathcal{V}^T$,

1) for all ζ_1^T s.t. $\sup_{\tilde{x}_1^T\in\mathcal{B}(f(x_1^T))}\gamma(\tilde{x}_1^T,\zeta_1^T)=1$: $P^{*,f}_{\zeta_1^T|X_1^T}(\zeta_1^T|x_1^T)$ satisfies

$$\sum_{\tilde{x}_1^T\in\mathcal{B}_f(x_1^T)} P^{*,f}_{X_1^T}(\tilde{x}_1^T)\sum_{\zeta_1^T} P^{*,f}_{\zeta_1^T|X_1^T}(\zeta_1^T|\tilde{x}_1^T)\sup_{\tilde{x}_1^T\in\mathcal{B}_f(x_1^T)}\gamma(\tilde{x}_1^T,\zeta_1^T) = \left(\sum_{\tilde{x}_1^T\in\mathcal{B}_f(x_1^T)} P^{*,f}_{X_1^T}(\tilde{x}_1^T)\right)\wedge\alpha.$$

2) $\forall\zeta_1^T$ s.t. $|\{x_1^T\in\mathcal{V}^T:\gamma(x_1^T,\zeta_1^T)=1\}|=0$: $P^{*,f}_{X_1^T,\zeta_1^T}(x_1^T,\zeta_1^T)$ satisfies

$$\sum_{\tilde{x}_1^T\in\mathcal{B}_f(x_1^T)} P^{*,f}_{X_1^T}(x_1^T)\sum_{\zeta_1^T:|\{x_1^T:\gamma(x_1^T,\zeta_1^T)=1\}|=0} P^{*,f}_{\zeta_1^T|X_1^T}(\zeta_1^T|x_1^T) = \left(\left(\sum_{\tilde{x}_1^T\in\mathcal{B}_f(x_1^T)} P^{*,f}_{X_1^T}(\tilde{x}_1^T)\right)-\alpha\right)_+.$$

3) all other cases of ζ_1^T: $P^{*,f}_{X_1^T,\zeta_1^T}(x_1^T,\zeta_1^T)=0$.

Proof of Theorem 8. When f is an identity mapping, it is equivalent to Theorem 2. When $f : \mathcal{V}^T \to [K]$ is some other function, following from the proof of Theorem 2, we consider three cases.

- **Case 1:** $\mathcal{S} \cap ([K] \cup \mathcal{V}^T) \neq \emptyset$ but $|\mathcal{S}| < K$. It is impossible for the detector to detect all the watermarked text sequences. That is, there exist $\tilde{x}_1^T$ such that for all ζ_1^T, $\gamma(\tilde{x}_1^T, \zeta_1^T) = 0$. Under this case, in Appendix M, $C(f(\tilde{x}_1^T)) = 0 \neq (\sum_{x_1^T : f(x_1^T) = f(\tilde{x}_1^T)} P_{X_1^T}(x_1^T)) \wedge \alpha$, which means the f-robust Type-II error cannot reach the lower bound.

- **Case 2:** $\mathcal{S} \cap ([K] \cup \mathcal{V}^T) \neq \emptyset$ but $|\mathcal{S}| = K$. Under this condition, the detector needs to accept the form $\gamma(X_1^T, \zeta_1^T) = \mathbb{1}\{f(X_1^T) = g(\zeta_1^T)\}$ so as to detect all possible watermarked text. Otherwise, it will degenerate to Case 1. We can see $f(X_1^T)$ as an input variable and rewrite the detector as $\gamma'(f(X_1^T), \zeta_1^T) = \gamma(X_1^T, \zeta_1^T) = \mathbb{1}\{f(X_1^T) = g(\zeta_1^T)\}$. Similar the proof technique of Theorem 2, it can be shown that $C(k)$ in Appendix M cannot equal $(\sum_{x_1^T : f(x_1^T) = k} P_{X_1^T}(x_1^T)) \wedge \alpha$ for all $k \in [K]$, while the worst-case f-robust Type-I error remains upper bounded by α for all $Q_{X_1^T}$ and ϵ.

- **Case 3:** Let $\Xi_\gamma(x_1^T) \coloneqq \{\zeta_1^T \in \mathcal{Z}^T : \gamma(x_1^T, \zeta_1^T) = 1\}$. $\exists x_1^T, y_1^T \in \mathcal{V}^T$, s.t. $f(x_1^T) \neq f(y_1^T)$ and $\Xi_\gamma(x_1^T) \cap \Xi_\gamma(y_1^T) \neq \emptyset$. For any detector $\gamma \notin \Gamma_f^*$ that does not belong to Cases 1 and 2, it belongs to Case 3. Let us start from a simple case where $T = 1$, $\mathcal{V} = \{x_1, x_2, x_3\}$, $K = 2$, $\mathcal{Z} = \{\zeta_1, \zeta_2, \zeta_3\}$, and $\mathcal{S} = [2]$. Consider the mapping f and the detector as follows: $f(x_1) = f(x_2) = 1$, $f(x_3) = 2$, $\gamma(x_1, \zeta_1) = \gamma(x_1, \zeta_1) = 1$, $\gamma(x_3, \zeta_2) = 1$, and $\gamma(x, \zeta) = 0$ for all other pairs (x, ζ). When $C(k) = (\sum_{x_1^T : f(x_1^T) = k} P_{X_1^T}(x_1^T)) \wedge \alpha$ for all $k \in [K]$, i.e.,

$$P_{X,\zeta}(x_1, \zeta_1) + P_{X,\zeta}(x_1, \zeta_2) + P_{X,\zeta}(x_2, \zeta_1) + P_{X,\zeta}(x_2, \zeta_2) = (P_X(x_1) + P_X(x_2)) \wedge \alpha,$$
$$\text{and} \quad P_{X,\zeta}(x_3, \zeta_2) = P_X(x_3) \wedge \alpha,$$

then the worst-case f-robust Type-I error is lower bounded by

$$\begin{aligned}
&\max_{Q_{X_1^T}} \mathbb{E}_{Q_{X_1^T} \otimes P_{\zeta_1^T}} \left[\sup_{\tilde{x}_1^T \in \mathcal{B}_f(X_1^T)} \mathbb{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 1\} \right] \\
&\geq \mathbb{E}_{P_{\zeta_1^T}} \left[\sup_{\tilde{x}_1^T \in \mathcal{B}(1)} \mathbb{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 1\} \right] \\
&= (P_X(x_1) + P_X(x_2)) \wedge \alpha + P_X(x_3) \wedge \alpha \\
&> \alpha, \quad \text{if } P_X(x_1) + P_X(x_2) > \alpha \text{ or } P_X(x_3) > \alpha.
\end{aligned}$$

Thus, for any Q_X such that $\{P_X \in \mathcal{P}(\mathcal{V}) : \mathsf{D}_{\mathsf{TV}}(P_X, Q_X) \leq \epsilon\} \subseteq \{P_X \in \mathcal{P}(\mathcal{V}) : P_X(x_1) + P_X(x_2) > \alpha \text{ or } P_X(x_2) > \alpha\}$, the false-alarm constraint is violated when $C(k) = (\sum_{x_1^T : f(x_1^T) = k} P_{X_1^T}(x_1^T)) \wedge \alpha$ for all $k \in [K]$. The result can be generalized to larger $(T, \mathcal{V}, \mathcal{Z}, K, \mathcal{S})$, other functions f and other detectors that belong to Case 3.

In conclusion, if and only if $\gamma \in \Gamma^*$, the minimum Type-II error attained from (Opt-R) reaches the universal minimum f-robust Type-II error $\beta_1^*(f, Q_{X_1^T}, \epsilon, \alpha)$ in (17) for all $Q_{X_1^T} \in \mathcal{P}(\mathcal{V}^T)$ and $\epsilon \in \mathbb{R}_{\geq 0}$.

Under the watermarking scheme $P^{*,f}_{X_1^T, \zeta_1^T}$, the f-robust Type-I and Type-II errors are given by:

f-robust Type-I error:

$$\begin{aligned}
&\because \forall y_1^T \in \mathcal{V}^T, \quad \mathbb{E}_{P^{*,f}_{\zeta_1^T}} \left[\sup_{\tilde{x}_1^T \in \mathcal{B}_f(y_1^T)} \mathbb{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 1\} \right] \\
&= \sum_{\zeta_1^T} \sum_{x_1^T} P^{*,f}_{X_1^T, \zeta_1^T}(x_1^T, \zeta_1^T) \sup_{\tilde{x}_1^T \in \mathcal{B}_f(y_1^T))} \mathbb{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 1\} \\
&= \sum_{x_1^T \in \mathcal{B}_f(y_1^T)} P^{*,f}_{X_1^T}(x_1^T) \sum_{\zeta_1^T} P^{*,f}_{\zeta_1^T | X_1^T}(\zeta_1^T | x_1^T) \sup_{\tilde{x}_1^T \in \mathcal{B}_f(y_1^T)} \mathbb{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 1\}
\end{aligned}$$

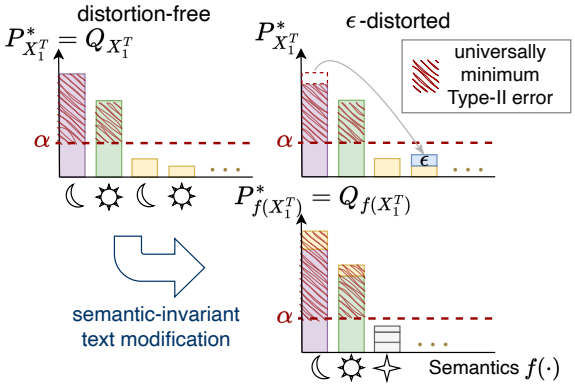

Figure 6: Universally minimum Type-II error w/o distortion and with semantic-invariant text modification.

$$= \left(\sum_{x_1^T \in \mathcal{B}_f(y_1^T)} P^{*,f}_{X_1^T}(x_1^T) \right) \wedge \alpha \leq \alpha,$$

and since any distribution $Q_{X_1^T}$ can be written as a linear combinations of $\delta_{y_1^T}$,

$$\therefore \sup_{Q_{X_1^T}} \mathbb{E}_{Q_{X_1^T} P^{*,f}_{\zeta_1^T}} \left[\sup_{\tilde{x}_1^T \in \mathcal{B}_f(X_1^T)} \mathbb{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 1\} \right] \leq \alpha.$$

f-robust Type-II error:

$$\begin{aligned}
&1 - \mathbb{E}_{P^{*,f}_{X_1^T,\zeta_1^T}} \left[\sup_{\tilde{x}_1^T \in \mathcal{B}_f(X_1^T)} \mathbb{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 1\} \right] \\
&= 1 - \sum_{x_1^T} \sum_{\zeta_1^T} P^{*,f}_{X_1^T,\zeta_1^T}(x_1^T, \zeta_1^T) \sup_{\tilde{x}_1^T \in \mathcal{B}_f(x_1^T)} \mathbb{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 1\} \\
&= 1 - \sum_{k \in [K]} \sum_{x_1^T \in \mathcal{B}(k)} P^{*,f}_{X_1^T}(x_1^T) \sum_{\zeta_1^T} P^{*,f}_{\zeta_1^T|X_1^T}(\zeta_1^T | x_1^T) \sup_{\tilde{x}_1^T \in \mathcal{B}(k)} \mathbb{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 1\} \\
&= 1 - \sum_{k \in [K]} \left(\left(\sum_{x_1^T \in \mathcal{B}(k)} P^{*,f}_{X_1^T}(x_1^T) \right) \wedge \alpha \right) \\
&= \sum_{k \in [K]} \left(\left(\sum_{x_1^T \in \mathcal{B}(k)} P^{*,f}_{X_1^T}(x_1^T) \right) - \alpha \right)_+ .
\end{aligned}$$

The optimality of $P^{*,f}_{X_1^T,\zeta_1^T}$ is thus proved. □

Figure 6 compares the universally minimum Type-II errors with and without semantic-invariant text modification.

O Implementation of Watermarking Scheme with Uniform P_{ζ_t}

O.1 Algorithm Description

Algorithm 3 describes the optimal watermarking scheme with uniform P_{ζ_t}. We first uniformly sample ζ_t from $\mathcal{Z} = \{h_{\mathsf{key}}(x)\}_{x \in \mathcal{V}}$. Then, with the sampled ζ_t, we can derive the new NTP distribution such that $P_{X_t|x_1^{t-1},u}(x) = |\mathcal{V}| \min\{Q_{X_t|x_1^{t-1},u}(x), \frac{1}{|\mathcal{V}|}\}$ for $h_{\mathsf{key}}(x) = \zeta_t$, while $P_{X_t|x_1^{t-1},u}(x) = \frac{|\mathcal{V}|\left(Q_{X_t|x_1^{t-1},u}(x) - \frac{1}{|\mathcal{V}|}\right)_+ \cdot \left(\frac{1}{|\mathcal{V}|} - Q_{X_t|x_1^{t-1},u}(h^{-1}_{\mathsf{key}}(\zeta_t))\right)_+}{\mathsf{D}_{\mathsf{TV}}(Q_{X_t|x_1^{t-1},u}, \mathrm{Unif}(\mathcal{V}))}$ otherwise. Next token is then sampled from obtained $P_{X_t|x_1^{t-1},u}(x)$.

Algorithm 4 outlines the corresponding detection process. For any given suspicious text, we analyze each token sequentially, mirroring the generation process. First, we uniformly sample ζ_t using previous tokens as a hash. Then, we compute the score as $\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{h_{\texttt{key}}(x_t) = \zeta_t\}$. Any text with a score greater than a threshold $\lambda \in (0, 1)$, will be classified as watermarked.

Algorithm 3 Watermarked Text Generation with Uniform P_{ζ_t}

Input: Language Model Q, Vocabulary $\mathcal{V}$, Prompt u, Secret $\texttt{key}$, Token-level False alarm η

1: $\mathcal{Z} \leftarrow \{h_{\texttt{key}}(x)\}_{x\in\mathcal{V}}$
2: **for** $t = 1, \ldots, T$ **do**
3: Compute a hash of previous n tokens, and use it as a seed to uniformly sample ζ_t from $\mathcal{Z}$.
4: $$P_{X_t|x_1^{t-1},u}(x) = \begin{cases} |\mathcal{V}| \min\{Q_{X_t|x_1^{t-1},u}(x), \frac{1}{|\mathcal{V}|}\}, & \text{if } h_{\texttt{key}}(x) = \zeta_t; \\ \frac{|\mathcal{V}|\left(Q_{X_t|x_1^{t-1},u}(x) - \frac{1}{|\mathcal{V}|}\right)_+ \cdot \left(\frac{1}{|\mathcal{V}|} - Q_{X_t|x_1^{t-1},u}(h_{\texttt{key}}^{-1}(\zeta_t))\right)_+}{\mathsf{D}_{\mathsf{TV}}(Q_{X_t|x_1^{t-1},u}, \mathrm{Unif}(\mathcal{V}))}, & \text{otherwise,} \end{cases}$$
5: Sample $x_t \sim P_{X_t|x_1^{t-1},u}$
6: **end for**

Output: Watermarked text $x_1^T = (x_1, \ldots, x_T)$.

Algorithm 4 Watermarked Text Detection with Uniform P_{ζ_t}

Input: Language Model Q, Vocabulary $\mathcal{V}$, Prompt u, Secret $\texttt{key}$, Token-level False alarm η

1: $\mathcal{Z} \leftarrow \{h_{\texttt{key}}(x)\}_{x\in\mathcal{V}}$
2: score = 0
3: **for** $t = 1, \ldots, T$ **do**
4: Compute a hash of previous n tokens, and use it as a seed to uniformly sample ζ_t from $\mathcal{V}$.
5: score = score $+\mathbb{1}\{h_{\texttt{key}}(x_t) = \zeta_t\}$
6: **if** score $> T\lambda$ **then**
7: **return** 1 ▷ Input text is watermarked
8: **else**
9: **return** 0 ▷ Input text is unwatermarked
10: **end if**
11: **end for**