Fast segmentation of watermarked texts from large language models through epidemic change-points framework

Soham Bonnerjee
Department of Statistics, University of Chicago, USA
and
Subhrajyoty Roy
Department of Statistics and Data Science, Washington University in St. Louis, USA
and
Sayar Karmakar
Department of Statistics, University of Florida, USA

December 4, 2025

Abstract

With the growing use of large language models, concerns over content authenticity have spurred a variety of watermarking schemes. These schemes use secret keys to detect machine-generated text while remaining imperceptible to readers. Detection typically reduces to statistical hypothesis testing for the presence of watermarks, a topic that is now well studied. In contrast, the finer-grained task of localizing which segments of a text are watermarked is much less explored; existing approaches often lack scalability or guarantees robust to paraphrasing and post-editing. We bring a new perspective to this segmentation problem through the lens of *epidemic change-points* and, by exploiting this connection, propose `WISER`, a novel and computationally efficient watermark segmentation algorithm. We establish finite-sample error bounds and consistency for detecting multiple watermarked segments in a single text. Complementing these theoretical results, our extensive numerical experiments show that `WISER` outperforms state-of-the-art baseline methods, both in terms of computational speed as well as accuracy, on various benchmark datasets embedded with diverse watermarking schemes. Together, these theoretical and empirical results position `WISER` as an effective tool for watermark localization and illustrate how classical statistical ideas can yield theoretically valid and computationally efficient solutions to a modern problem of immediate importance.

Keywords: Watermarking, Large language model, Change-point, Epidemic change-point

1 Introduction

Recent years have seen widespread popularity and adoption of Large Language Models (LLM) in areas such as media, education, healthcare, and finance- domains where content creation, ownership and automation (Touvron et al. 2023, Achiam et al. 2023) occupies central importance. However, an unfortunate consequence of the exponential ascent of LLMs has been an increased propagation of synthetic texts across the internet. This has raised significant security and legal concerns regarding privacy, content authenticity and copyright infringement over multiple domains (Megías et al. 2022, Bender et al. 2021, Crothers et al. 2023, Liang et al. 2024, Milano et al. 2023, Radford et al. 2023, Chen & Shu 2023, Woodcock 2023). In particular, the ability of LLMs to generate a large volume of texts makes them vulnerable to intended or unintended misuse by entities, often in violation of the governing guidelines to achieve potential plagiarism or deceit (Ahmed et al. 2021, Lee et al. 2023). For example, recently, the use of LLM-generated text without proper attribution has evolved into a full-fledged quagmire in the lawsuit between New York Times and OpenAI (Grynbaum & Mac 2023). In the same mold, our colleagues in academia, and educators more generally, often face a perhaps legally less challenging but equally important issue: AI-assisted education. The use of AI may, prima facie, be encouraged in many low-stakes situations. However, an increased proliferation of LLM-generated texts in critical assessments not only constitutes a malpractice, but also deprives students of the potential to embark upon an important learning curve by themselves, while simultaneously propagating unfair advantages to more privileged students who have access to newer LLM models (Milano et al. 2023, Wang et al. 2024, Darvishi et al. 2024).

Such concerns were initially addressed by attempting to identify LLM-generated texts via specific patterns or properties of the said texts, such as cross-entropy or perplexity (Mitchell et al. 2023, ZeroGPT 2024, Radvand et al. 2025). However, the shortcomings of this approach have become increasingly evident as more and more language models gain the ability to mimic the

quirks of a human-generated text. Open-access, publicly funded large language models have been conceptualized as another alternative, mitigating strategy (Akiki et al. 2022, Workshop et al. 2022, Shrestha et al. 2023, Li et al. 2023, Üstün et al. 2024). In a different direction, and probably most relevant with regards to fraud detection in education, "Watermarking methods" have been proposed (Christ et al. 2024, Aaronson 2023), and widely adopted (Biden 2023, Bartz & Hu 2023) as a detection mechanism. Watermarking schemes primarily exploit the tokenization structure of large language models. In principle, given a sequence of tokens $\omega_1 \dots \omega_{t-1}$, the LLM generates ω_t from a multinomial distribution P_t over the dictionary $\mathcal{W}$, where P_t, the *Next Token Distribution* (NTP) is allowed to depend on previous tokens $\omega_1, \dots, \omega_{t-1}$. Then, watermarking is used to embed statistical signals into LLM-generated tokens, which remain largely unnoticeable without additional information. The key insight behind watermark-based detection schemes is the use of the underlying randomness of LLM-generated outputs by incorporating pseudo-randomness into the text-generation process. When a third-party user publishes text potentially containing LLM-generated outputs with watermarks, the coupling between the LLM-generated text and the pseudo-random numbers serves as a signal that can be used for detecting the watermark. Crucially, the properties of watermarks allow the user to detect machine-generated texts without requiring knowledge of any particular properties of the text or the LLM. For example, it is conceivable that the academic institution penalizing LLM-generated texts may gain access to the pseudo-random numbers from the particular LLM they deploy in their network system used by the students. We emphasize that the knowledge of these pseudo-random numbers is imperative for the detection mechanism to work, making the effect of watermarking untraceable to general users, who usually do not have access to such "keys".

This usefulness has stimulated a plethora of research proposing myriad watermarking schemes (Kirchenbauer et al. 2024, Fernandez et al. 2023, Golowich & Moitra 2024, Hu et al. 2024, Wu et al. 2024, Zhao et al. 2025, Zhao, Ananth, Li & Wang 2024, Liu & Bu 2024, Zhu et al. 2024). Concurrently, much attention has landed on the pursuit of efficient, statistically valid detection

schemes (Li, Ruan, Wang, Long & Su 2025*b*, Kuditipudi et al. 2024, Cai et al. 2024, Huang et al. 2023, Li, Ruan, Wang, Long & Su 2025*a*, Cai et al. 2025), as well as on the more general problems of machine-generated text detection or model equality testing (Lavergne et al. 2008, Solaiman et al. 2019, Gehrmann et al. 2019, Su et al. 2023, Mitchell et al. 2023, Huang et al. 2023, Vasilatos et al. 2023, Hans et al. 2024, Li, Ruan, Wang, Long & Su 2025*b*, Kuditipudi et al. 2024, Cai et al. 2024, Gao et al. 2025, Song et al. 2025, Radvand et al. 2025). These detection schemes usually rely on the knowledge of the pseudo-random keys or deterministic hash functions to perform a composite-vs-composite test of hypotheses: H_0 : the entire text $\omega_1 \dots \omega_n$ is un-watermarked (i.e. human generated), vs H_1 : the entire text is watermarked or H_1' : the text contains watermarked segments (Mitchell et al. 2023, Bao et al. 2024, Li, Ruan, Wang, Long & Su 2025*b*, Zhou et al. 2025). Usually, such tests depend on the *pivot statistic* Y_t s, which are formed from the token ω_t and the watermarking keys ζ_t. The virtues of the pivot statistics stem from their ancillarity with respect to the next token distributions $\{P_t\}$, allowing it to be used without requiring specific knowledge about the LLM architecture or its NTP distributions. Recent advances in this direction have started to shed light on detecting more sophisticated modifications of watermarking by allowing arbitrary data misappropriation Cai et al. (2025) and arbitrary modifications such as deletion and replacements Li, Ruan, Wang, Long & Su (2025*a*), Xie et al. (2025). However, somewhat curiously, the relatively harder and more fine-grained problem of precisely localizing the watermarked segments from an input text has received only sparse attention. Apart from WinMax (Kirchenbauer et al. 2024), which focuses only on Red-Green watermarking, to the best of our knowledge, the only algorithms tackling the segmentation problem in its generality are Li et al. (2024), Pan et al. (2025) and Zhao, Liao, Wang & Li (2024). Most of these algorithms are prohibitively slow and thus unsuited for long texts. Moreover, to the best of our knowledge, no such algorithm designed to efficiently identify multiple watermarked segments has sufficient theoretical validity. This gap in the literature is also pointed out by Li, Wen, He, Wu, Long & Su (2025).

In this paper, we propose `WISER` (**W**atermark **I**dentification via **S**egmenting **E**pidemic **R**egions): a *first-of-its-kind* computationally efficient and provably consistent algorithm to locate multiple watermarked segments from mixed-source input texts. Our method is inspired from the classical notion of *epidemic* change-points; this perspective is instrumental for both the theoretical validity and computational efficiency of our algorithm. We summarize our main contributions as follows.

Firstly, in §2, we introduce a novel, *epidemic change-point* perspective on the watermark segmentation problem by exploiting an inherent property of the watermarking schemes. In particular, research dealing with testing for the existence of watermarks essentially hinges upon a score function h applied over the pivot statistics Y_t, which usually has the property that $\mathbb{E}[h(Y_t)]$ is much larger for the watermarked tokens than for un-watermarked ones. This property can be visualized in Figure 1, and is elaborated on with examples in Section 2.1. While this *elevated alternatives property* (Assumption 2.2) is crucial in achieving significant power for the testing problems, it has not been formally described and analyzed in this context. However, for a localization problem, this property readily relates it to a separate classical problem of *epidemic* change-point detection. Roughly speaking, an epidemic change-point refers to a situation where a stochastic process deviates in one of its features in an interval and returns to the baseline. In simple words, the changes in the related features occur in interval patches, and outside these patches the process behaves in an i.i.d. or stationary fashion. Since the score-appended pivot-statistics $h(X_t)$ exhibit very similar behavior in watermarked tokens, this interpretation of patches as epidemic change-point intervals enables us to re-purpose some of the classical insights of change-point literature into a state-of-the-art algorithm to provably locate them and thus solve a modern problem in the area of AI-moderation. Even though Li et al. (2024), Li, Liu & Li (2025) also relate the watermark segmentation problem to a change-point localization problem, their insights are rather limited, since they identify the end-points of watermarked patches as distinct change-points, which does not respect the nature of watermarked tokens appearing as intervals.

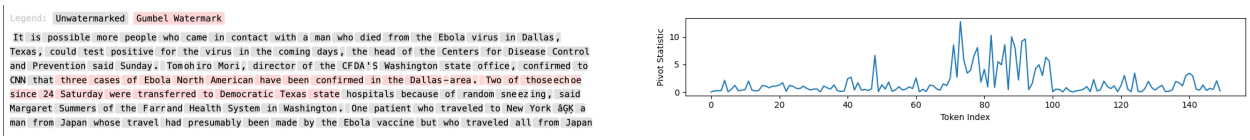

Figure 1: (Left) A text with watermarked tokens 70-100. (Right) The corresponding plot of pivot statistics vs. token.

Secondly, in §4, we transform the epidemic change-point insight into a tractable algorithm, tuned to the peculiarities of the watermarking framework. Specifically, in contrast to the usual setting of independent or stationary random variables in the epidemic change-point literature (Levin & Kline 1985, Hušková & Slabý 1995, Chen et al. 2016), we work with a highly non-stationary setting, devoid of any direct regularity assumption for watermarked-tokens. Motivated by several studies on irregular change-point analysis (Kley et al. 2024), we devise the `WISER` algorithm, which is valid irrespective of the LLMs or NTPs. The algorithm is illustrated schematically in Figure 2, and also discussed in the Appendix §A. In principle, our algorithm is simple to describe. The *epidemic* interpretation produces a natural estimate for the case of a single watermarked segment, and the general case of multiple watermarked segments can then be dealt with by appropriately restricting the search spaces for each of these segments. The number of such segments is estimated by a series of carefully orchestrated steps (such as block-based tests, and a threshold-based deletion of false-positive blocks), and we further restrict the search space is ensured to reduce the computational burden. To summarize, our algorithm simply works with the pivot statistics and the elevated alternatives property, and brings insights from the epidemic change-point theory to tackle the potentially arbitrary non-stationary dependence typically displayed by the pivot statistics corresponding to the watermarked tokens.

Thirdly, in §4 we rigorously establish the theoretical validity of our algorithm in very general scenarios. The theoretical validity of the `WISER` segmentation algorithm arises as an automatic consequence of our perspective. Additionally, we motivate the local estimate used in the last

stage of WISER by proving in Theorem 3.1 that it is consistent in the single watermarked-segment case. To the best of our knowledge, WISER is the *first watermark segmentation algorithm with complete theoretical guarantees in the most general case.* It is important to note that the regime of non-stationarity in one or more epidemic patches is different from the usual multiple change-point regime with an even number of breakpoints due to how we perceive signal and noise in a statistical detection problem. Moreover, as we already mentioned above, there is inherent irregularity intrinsic owing to how watermarked texts are generated. Our theoretical results settle these issues in a comprehensive fashion. Part of our proof techniques are based on moment and cumulant generating functions, as well as Danskin (1967)'s results, which are, novel to both change-point or watermark literature to the best of our knowledge and these tools can be of independent interest.

Finally, the ingenuity of our algorithm lies not only in its amalgamation of different ideas from statistics, but also in its practicality. In the numerical experiments §5-6, the theoretical guarantees are reflected in WISER's superiority over other competitive methods across different watermarking schemes and different language models. In the Appendix §C, we provide additional and extensive numerical experiments to further reinforce the effectiveness of our algorithm, as well as highlighting the novelty of our algorithm compared to the other algorithms in the literature. Another key aspect of its enhanced performance is its speed. WISER is specifically designed with many localized steps that reduce its run-time, thereby making it, to the best of our knowledge, the only $O(n)$ watermark segmentation algorithm with provable theoretical guarantees.

1.1 Notations

We delineate some of the notations to be used throughout this paper. The set $\{1, \ldots, n\}$ is denoted by $[n]$. The d-dimensional Euclidean space is $\mathbb{R}^d$. For a vector $a \in \mathbb{R}^d$, $|a|$ denotes its Euclidean norm. For a random vector $X \in \mathbb{R}^d$, we denote $\|X\| := \sqrt{\mathbb{E}[|X|^2]}$. Throughout the paper, we use the usual Landau notation $O(\cdot), o(\cdot)$ for sequences of real numbers. The analogous stochastic

versions, corresponding to stochastic boundedness and in-probability, convergence, are denoted by $O_{\mathbb{P}}(\cdot)$ and $o_{\mathbb{P}}(\cdot)$ respectively. We also write $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for some constant $C > 0$, and $a_n \asymp b_n$ if $C_1 b_n \leq a_n \leq C_2 b_n$ for some constants $C_1, C_2 > 0$. Finally, $\mathcal{L}(X)$ denotes the law of X.

2 Watermark segmentation: epidemic change-point perspective

Before we introduce our novel perspective in the context of locating watermarked segments, it is instrumental to establish a consistent framework of watermarking in LLM-generated texts. Let $\mathcal{W}$ denote the dictionary, enumerated as $1, 2, \ldots, |\mathcal{W}|$. Given a text input in a tokenized form $\omega_1 \ldots \omega_{t-1}$, a watermarked LLM generates the next token ω_t in an autoregressive manner as $\omega_t = S(P_t, \zeta_t)$, where $P_t = (P_{t,w})_{w=1}^{|\mathcal{W}|}$ is the next token probability (NTP) distribution at step t; S is a deterministic decoder function, and ζ_t is the pseudo-random variable at t. We grant Assumption 2.1 for the ζ_t's.

Assumption 2.1. *For any text $\omega_{1:n}$, there exists corresponding pseudo-random variables $\zeta_{1:n}$ available to the verifier, such that if the token ω_t at step t is un-watermarked, then ω_t and ζ_t are independent conditional on $\omega_{1:(t-1)}$.*

It may seem that this assumption invalidates human edits after the LLM generates a text. However, in Appendix §B, we discuss how Assumption 2.1 applies to the mixed-source texts allowing for human edits.

2.1 Pivot statistics and elevated alternatives

Note that, a text $\omega_{1:n}$ with K disjoint watermarked intervals $I_1, \ldots, I_K$, $I_j \subset [n]$ for $j \in [K]$, can be modeled as

$$w_t \sim \begin{cases} P_t, & t \notin I_0 := \cup_{l=1}^{K} I_k, \\ S(P_t, \zeta_t), & \text{otherwise}, \end{cases} \qquad t = 1, 2, \ldots, n. \tag{1}$$

We are interested in the statistical problem of estimating the individual intervals $I_1, \ldots, I_K$ as well as K. Before proceeding further, it is appropriate to formally introduce the notion of pivot statistics.

Definition 2.1. *$Y(\omega, \zeta)$ is called a pivot statistic if $\mathcal{L}(Y)$ is same for all $\omega \in \mathcal{W}$.*

Pivot statistic has been extremely effective in providing statistically valid testing strategies for the existence of watermarks in mixed-source texts (Li, Ruan, Wang, Long & Su 2025*b*,*a*, Cai et al. 2024), however, in what follows, we will demonstrate their effectiveness in aiding a localization algorithm. This effectiveness is a result of a simple property of the pivot statistics; they metamorphose the conditional independence of ω_t and ζ_t for un-watermarked tokens into P_t-independent distributions. Formally, this property is described in the following result.

Lemma 2.2. *If S denotes the set of un-watermarked tokens, then $\{Y_t\}_{t \in S}$ are i.i.d.*

This ancillarity is heavily used in all the available statistical analysis of watermarked schemes; nevertheless, for the sake of completion we provide a proof in Appendix §D.3. Lemma 2.2 enables us to use the notation $\mu_0 := \mathbb{E}_0[Y(\omega, \zeta)]$ as the expectation of the pivot statistic Y when the token $\omega \sim P$ is not watermarked; on the other hand, $\mathbb{E}_{1,P}[Y(\omega, \zeta)]$ will denote expectation with respect to the randomness of ζ (i.e. conditional on P) when ω is watermarked according to (S, ζ)-mechanism. Finally, we denote $Y_t := Y_t(\omega_t, \zeta_t)$. Note that since Y_t is a pivot statistic, so is $h(Y_t)$ for any *score* function $h : \mathbb{R} \to \mathbb{R}$. Usual tests for watermark detection look at $\sum_{t=1}^{n} h(Y_t)$ as a statistic for a one-sided test, and put considerable effort into constructing an effective score function h (Kirchenbauer et al. 2024, Zhao, Liao, Wang & Li 2024, Li, Ruan, Wang, Long & Su 2025*b*, Cai et al. 2025). Intrinsic to this construction, even though never explicitly stated, is the assumption that $\mathbb{E}_{1,P}[h(Y)]$ is usually larger than μ_0 for any possible NTP distribution P. This hypothesis of "elevated alternatives" can also be empirically viewed in Figure 1.

We formalize this observation with the following hypothesis.

Assumption 2.2 (Elevated Alternatives Hypothesis)**.** *Assume that the next token distribution (NTP) P belongs to a distribution class $\mathcal{P}$. Then, there exists $d > 0$ such that $\inf_{P\in\mathcal{P}} \mathbb{E}_{1,P}[h(Y)] \geq \mu_0 + d$, where $\mathbb{E}_{1,P}(\cdot) = \mathbb{E}_1[\cdot|P]$ denotes the unknown distribution of $h(Y)$ when watermarking is implemented on the NTP $P \in \mathcal{P}$ via $S(P, \cdot)$.*

This assumption entails that the pivot statistics is effective conditional on any possible NTP from the class $\mathcal{P}$, ruling out trivial cases such as $Y(\omega, \zeta) \equiv \zeta$. Most standard watermarking schemes satisfy Assumption 2.2; see the following for some concrete examples.

2.1.1 Examples to Assumption 2.2

In this section, we justify the elevated alternative hypothesis Assumption 2.2 by illustrating its occurrence through two popular watermarking schemes.

Example (Gumbel Watermark, Aaronson (2023)). Let $\zeta = (U_w)_{w\in\mathcal{W}}$ consist of $|\mathcal{W}|$ i.i.d. copies of $U(0, 1)$. The Gumbel watermark is implemented as:

$$S^{\text{gum}}(\zeta, P) := \arg\max_{w\in\mathcal{W}} \frac{\log U_w}{P_w}, \tag{2}$$

The pivot statistic is taken as $Y_t = U_{t,\omega_t}$, $t \in [n]$. From Proposition 1 in Appendix, when $\Delta = 1/2$, $\inf_{P\in\mathcal{P}_\Delta} \mathbb{E}_{1,P}[h(Y)] \geq \sum_{n=1}^{\infty}(\frac{1}{n} - \frac{1}{n+2})$, which, in light of $h(Y) \sim \text{Exp}(1)$ entails that $d \geq 1/2$.

Example (Inverse Transform Watermark, Kuditipudi et al. (2024)). Consider an NTP distribution P and a permutation $\pi : \mathcal{W} \mapsto S_{|\mathcal{W}|}$, where $S_{|\mathcal{W}|}$ is the group of permutations of $\{1, 2, \ldots, |\mathcal{W}|\}$. Further consider the multinomial distribution $\{P_{\pi^{-1}(w)}\}_{w=1}^{|\mathcal{W}|}$. The CDF of this distribution takes the form

$$F(x; \pi) = \sum_{w'\in\mathcal{W}} P_{w'} \cdot \mathbf{1}_{\{\pi(w')\leq x\}}.$$

Taking as input $U \sim U(0,1)$, the generalized inverse of this CDF is defined as

$$F^{-1}(U;\pi) = \min\Big\{i : \sum_{w' \in \mathcal{W}} P_{w'} \cdot \mathbf{1}_{\{\pi(w') \le i\}} \ \geq \ U\Big\},$$

which, under the H_0 of no watermark, follows the multinomial distribution P after applying the permutation π. The inverse transform watermark is defined as the decoder:

$$\mathcal{S}^{\text{inv}}(P, \zeta) := \pi^{-1}\Big(F^{-1}(U;\pi)\Big).$$

Lemma 4.1 of Li, Ruan, Wang, Long & Su (2025*b*) indicates that under the alternative, the distribution of S^{inv} is intricately inter-related with the NTP P. To make the verification of Assumption 2.2 tractable, we impose a few assumptions. Assume $|\mathcal{W}| \to \infty$, and with $P_{t,(i)}$ denoting the i-th largest co-ordinate of the probability vector $P_{t,(i)}$ for every token t and $i \in [|\mathcal{W}|]$, we also assume

$$\lim_{|\mathcal{W}| \to \infty} P_{t,(1)} = 1 - \Delta \ \text{ and } \ \lim_{|\mathcal{W}| \to \infty} \log|\mathcal{W}| \cdot P_{t,(2)} = 0.$$

Consider the pivot statistic

$$Y_t = \Big|U_t - \eta(\pi_t(w_t))\Big|, \ \eta(i) := \frac{i-1}{|\mathcal{W}|-1}.$$

Under Theorem 4.1 of Li, Ruan, Wang, Long & Su (2025*b*), $\mathbb{E}_1[1-Y] = \frac{2+\Delta}{3}$, and $\mathbb{E}_0[1-Y] = \frac{2}{3}$. Therefore, here $d = \frac{\Delta}{3}$.

To summarize, the pivot statistics Y_t has a mean level μ_0 when the token ω_t is un-watermarked; on the other hand, we expect the pivot statistics to take comparatively larger values inside the watermarked segments. Interestingly, this observation establishes a ready-made connection to the notion of "epidemic change-points", sporadically explored in the classical time-series literature for

the past few decades. We discuss this novel perspective in the following section.

2.2 Watermarked interval in the context of epidemic change-point

We start with an epidemic changepoint model with a single change. The simplest and yet the most popular formulation of a 'mean-shift' epidemic model is as follows. Consider the time-series $X_i = \mu_i + Z_i$, where Z_i is mean-zero stationary process and

$$\mu_i = \mu \text{ if } i \in \{1, \cdots, p\} \cup \{q+1, \cdots, n\} \text{ and } \mu_i = \mu + d \text{ if } i \in \{p+1, \cdots, q\} \tag{3}$$

With K many true patches, this model reads as follows. For $1 < p_1 < q_1 < p_2 \cdots < q_k < n$,

$$\mu_i = \begin{cases} \mu + d_k, & i \in \{p_k + 1, \ldots, q_k\} \text{ for some } k = 1, \ldots, K, \\ \mu, & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n. \tag{4}$$

Epidemic changepoint is not new by any means. This framework originated with Levin & Kline (1985), who studied the testing for the existence of such epidemic patches for epidemiology applications, with a more comprehensive discussion in Yao (1993), Inclán & Tiao (1994). Later on, Hušková (1995), Csörgő & Horváth (1997), Chen et al. (2016) have discussed consistency, asymptotic theory as well as statistical powers of these epidemic estimators and accompanying tests. Other related papers discussing inference tailored to epidemic alternatives can be found in Račkauskas & Suquet (2004, 2006), Ning et al. (2012). Compared to the vast literature for usual change-point analysis, the epidemic change-point literature has been quite sparse, and even then, the focus has remained mostly on testing for the existence of such temporary departure rather than on locating these patches with provable statistical guarantees. In particular, the testing problem deals with the case $d_1 = d_2 = \cdots d_K = 0$. On the other hand, our work concerns simultaneously estimating the number of true locations K and the corresponding patches (p_i, q_i). The literature

on localizing multiple epidemic patches is even sparse (Zhao & Yau 2021, Juodakis & Marsland 2023), and seems to focus only on the much restricted setting on independent Gaussian observations. Moreover, as discussed in the Introduction as well, due to the nature of pivot statistics, we suffer from a certain irregularity induced by the non-stationarity in mean of the pivot-statistics for watermarked tokens. Therefore, any potential results or algorithms that might be obtained pertaining to model (3) or (4), are not directly applicable here. Instead, invoking Assumption 2.2, we can only assume that the means of the pivot statistics are separated from the null by at least some margin. This puts us in a position to solve an epidemic mean-shift problem of a new kind, where we can solve the case of localizing multiple patches accounting for this non-stationary departure in the mean of the pivot statistic.

Very recently Kley et al. (2024) proposed usual change-point detection under the presence of such irregular signals. Concretely, for noisy data of the form $X_t = \mu_t + Z_t,\ t = 1, \ldots, n$ where μ_t are means or signals and $(Z_t)_{t\in\mathbb{Z}}$ is a stationary mean-zero noise, they considered the following hypothesis testing problem with irregular 'non-constant-mean' alternative:

$$H_0 : \mu_1 = \cdots = \mu_n \text{ vs. } H_1 : \exists\, \tau \in \{2, \ldots, n\},\ d > 0 :\ \mu_1 = \cdots = \mu_{\tau-1}, \quad \mu_\tau, \ldots, \mu_n \geq \mu_1 + d.$$

They also proposed an estimation procedure for the location parameter τ. In this work, we extend their estimators to the epidemic alternative with properties dictated by Assumption 2.2, and provide guarantees of accurate localization. The analysis in Kley et al. (2024) is restricted to single change-point whereas the scenario of multiple patches with irregular signal within it comes naturally in our context. Moreover, the intrinsic dependence introduced by the context of how an LLM token sequence is generated also makes our premise for the error specification quite novel and thus brings out significant technical challenges. To address these challenges, we begin with a simpler problem segmenting of only one watermarked patch in §3.

3 Single watermarked patch

In this section, we underlay the development of our algorithm by starting with the simpler case of localizing a single watermarked patch. In particular, we propose an estimator to localize a single watermarked segment inside a text, and establish its theoretical consistency with finite sample results. Building on this estimate, in §4 we will formally propose the `WISER` algorithm to detect multiple patches.

We work with the pivot statistics $X_t = h(Y_t)$. Recall Lemma 2.2, the notation $\mu_0 = \mathbb{E}_0 X_t$, and Assumption 2.2. The pivot statistics are constructed so that under unwatermarked tokens, they behave like i.i.d. observations with a stable null mean μ_0. Under watermarking, however, the mean of X_t inside the true interval I_0 is not assumed constant: token-by-token perturbations can make it vary arbitrarily, and all we rely on is an elevated–alternatives condition, as Assumption 2.2 describes, $\inf_{P\in\mathcal{P}} \mathbb{E}_{1,P}[h(Y)] \geq \mu_0 + d$. Because of this irregularity, classical epidemic/CUSUM-type scans that presume a constant shift on the affected block are not directly applicable. Instead, we flip the viewpoint and search for an interval whose removal makes the remaining data look as close as possible to the null; this leads us to an initial interval estimator defined by minimizing a biased outside-of-interval surplus.

In this spirit, let $\tilde{d}$ be such that there exists $\rho \in (0,1)$ satisfying $d > 2\rho\tilde{d}$. Based on our discussion above, we adapt the estimator from Kley et al. (2024) for our particular 'epidemic' setting.

$$\hat{I} = \arg\min_{s,t\in[n]} \sum_{k\notin[s,t]} (X_k - \mu_0 - \rho\tilde{d}). \tag{5}$$

The role of the bias term $\rho\tilde{d}$ is crucial. By subtracting a positive buffer, we make each null token outside any candidate interval contribute a negative expected amount $-\rho\tilde{d}$, while any missed watermarked token left outside contributes at least $d - \rho\tilde{d}$, which is positive when the signal

dominates the buffer. Consequently, underestimating the interval leaves elevated points outside and increases the objective, whereas overestimating it removes extra null points and loses many negative contributions, also increasing the objective; the minimizer is therefore driven toward the smallest interval that excises all elevated tokens. Since the true elevation d is unknown, we use a proxy $\tilde{d}$ together with a tuning factor $\rho \in (0, 1)$ to remain conservative: $\tilde{d}$ provides a scale for the elevation we expect, and ρ controls the tradeoff between being too permissive (small ρ, risking overestimation from null fluctuations) and too strict (large ρ, risking loss of separation if d is not sufficiently bigger than $\rho\tilde{d}$).

The following theorem analyzes its convergence properties for the case of a single, uninterrupted watermarked region. Subsequently, we discuss some of its connotations in successive remarks.

Theorem 3.1. *Let* $\{X_t\}_{t=1}^n := \{h(Y_t)\}_{t=1}^n$ *be the pivot statistics based on the given input text, and assume that* $I_0 \subset \{1, \ldots, n\}$ *is the only watermarked interval. Grant Assumption 2.2. Denote*

$$\varepsilon_t = \begin{cases} X_t - \mu_0, t \notin I_0, \\ X_t - \mu_t,\ \mu_t := \mathbb{E}_{1,P_t}[X_t], t \in I_0. \end{cases}$$

Suppose the class of distributions $\mathcal{P}$ *is closed and compact, and there exists* $\eta > 0$ *such that* $\sup_{P\in\mathcal{P}} \mathbb{E}_{1,P}[\exp(\eta|\varepsilon|)] < \infty$. *Moreover, assume that* $\min\{\mathrm{Var}_0(\varepsilon), \sup_P \mathrm{Var}_{1,P}(\varepsilon)\} > 0$. *Consider the estimate (5) with* ρ *and* $\tilde{d}$ *satisfying* $d > 2\rho\tilde{d}$. *If there exists a constant* $c > 0$ *such that* $d \geq c$, *then* $|\hat{I}\Delta I_0| = O_{\mathbb{P}}((\rho\tilde{d})^{-1})$. *Here* Δ *is the symmetric difference operator and* $O_{\mathbb{P}}$ *hides constants independent of* $n, \tilde{d}, \rho$, *and* μ_0.

The $O(\tilde{(}\rho\tilde{d})^{-1})$ rate can further be sharpened to $O(\tilde{(}\rho\tilde{d})^{-2})$ under a local sub-Gaussianity condition (see Proposition D.2 in the Appendix §D). In fact, under very mild conditions, Theorem 3.1 already tackles a more general scenario compared to the only other theoretical result available in a similar context (Li et al. 2024). In contrast to a general watermarked patch, Li et al. (2024) considered

a specialized scenario, where only the first half of the text till an arbitrary point is watermarked, reducing the problem to a classical change-point setting.

The parameter $\tilde{d}$ serves as the *signal strength* in the convergence diagnostics of $\hat{I}$. It allows $\hat{I}$ to look for intervals such that the $\tilde{d}$-biased mean outside that interval is minimized. However, due to the restriction $d > 2\rho\tilde{d}$, since the minimum separation d in Assumption 4.1 is typically unknown, it cannot be used directly. In most cases (see examples in §2.1.1), a distribution-dependent lower bound $d_{\mathrm{L}} \leq d$ may be available, but relying on $\tilde{d} = d_{\mathrm{L}}$ often sacrifices power, as $\inf_{t\in[n]} \mathbb{E}_{1,P_t}[X_t - \mu_0]$ is usually much larger. Thus, a key step in practice is a data-driven yet valid choice of $\tilde{d}$, which we discuss in §4. The tuning parameter ρ adjusts the impact of $\tilde{d}$ and mitigates small errors in its selection. Choosing $\rho \approx 0$ is undesirable, as it causes $\hat{I}$ to overestimate I due to fluctuations above μ_0 under the null. Conversely, setting $\rho \approx 1$ can violate the requirement $d > 2\rho\tilde{d}$ when $\tilde{d}$ is large. Empirically, $\rho \in [0.1, 0.5]$ provides robust performance, and we revisit these choices in our discussion of `WISER` as well as the ablation studies in Appendix §C.3.

Remark 3.1 (Connection with other performance metric). Even though Theorem 3.1 controls the estimation error in terms of symmetric difference between estimated and true watermarked patches $\hat{I}$ and I respectively, it is straightforward to transform this result in terms of the more familiar Intersection-Over-Union metric $\mathrm{IOU}(I, \hat{I}) = |I \cap \hat{I}|/|I \cup \hat{I}|$ as $1 - \mathrm{IOU}(I, \hat{I}) = \frac{|I \Delta \hat{I}|}{|I \cup \hat{I}|} = O_{\mathbb{P}}\left(\frac{1}{|I|\rho\tilde{d}}\right)$. As the text size increases ($n \to \infty$), if $|I| = O(1)$, then the number of un-watermarked tokens is too large, overpowering the signal from the watermarked tokens. Under this "heavy-edit" regime, no non-trivial test statistic can differentiate between H_0 : the entire text $\omega_{1:n}$ is un-watermarked (i.e., human-generated) and H_1 : the entire text $\omega_{1:n}$ is watermarked, with reasonable power (Li, Wen, He, Wu, Long & Su 2025). The estimation being a harder problem than testing, it is therefore reasonable to assume $|I| \to \infty$ as $n \to \infty$. Therefore, Theorem 3.1 essentially entails that $\mathrm{IOU}(I, \hat{I}) \to 1$ as $n \to \infty$.

Despite the attractive theoretical properties of $\hat{I}$ given in (5), notwithstanding the yet unclear choice of $\tilde{d}$, there are a couple of practical roadblocks to deploying $\hat{I}$. Firstly, $\hat{I}$ has a computational complexity of $O(n^2)$, which is quite prohibitive for a large body of text one usually encounters. Secondly, it is not straightforward as to how $\hat{I}$ can be generalized to localize multiple watermarked segments. We answer these questions by proposing our `WISER` algorithm.

4 `WISER` segmenting multiple watermarked patches

The main motivation behind our proposed algorithm `WISER` is to use the estimator $\hat{I}$ on localized disjoint intervals that are more-or-less guaranteed to contain the true watermarked segments. Such intervals with guarantees are usually recovered as a consequence of some first-stage screening. For the convenience of readers, a schematic diagram of `WISER` containing the key steps, is illustrated in Figure 2. The detailed algorithm can be found in Appendix §A.

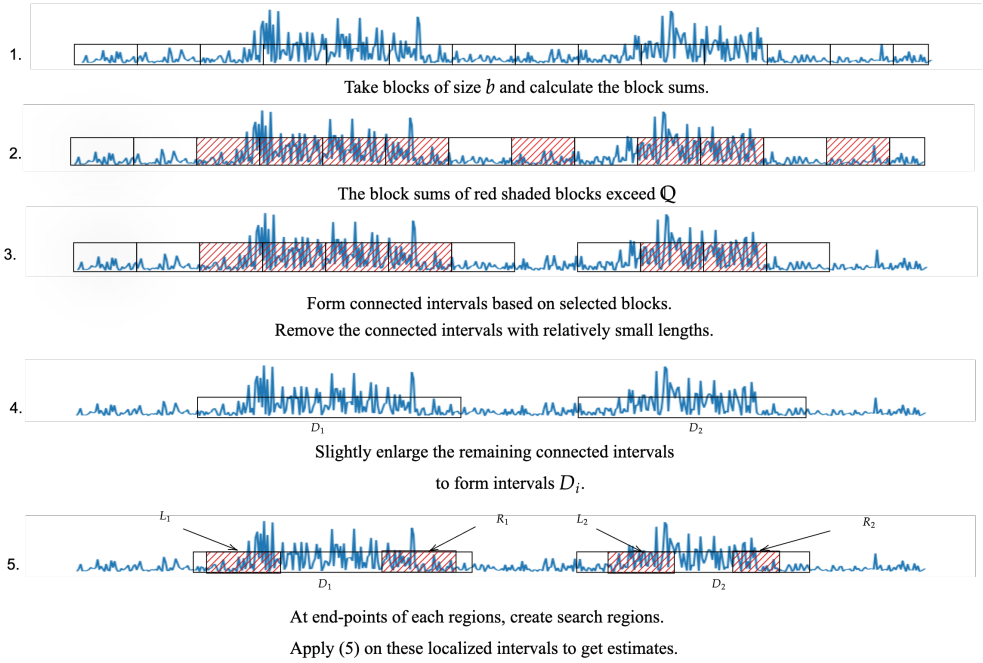

Figure 2: `WISER` in action with key steps.

Subsequently, we make a mild assumption that the true watermarked segments have a minimum

length, and are also well-separated to be considered as distinct segments. Formally, for two disjoint intervals $I_1 = (I_{1,L}, I_{1,R})$ and $I_2 = (I_{2,L}, I_{2,R})$, let $d(I_1, I_2) := \min\{|I_{1,L} - I_{2,R}|, |I_{1,R} - I_{2,L}|\}$.

Assumption 4.1 (Minimum separation)**.** *Let K be the number of true watermarked segments, with the segments themselves denoted by $I_j, j \in [K]$. Then there exists a constant $C_0 > 0$, such that $\min_{k\in[K]}\{|I_k| \wedge d(I_k, I_{k-1})\} \geq C_0 n^{\upsilon} \log n$ for some $\upsilon > 0$.*

Remark 4.1. In most practical scenarios, where a test for the existence of the watermark has sufficient power, the size of the watermarked patches will be significant, or should have high entropy. In fact, most of the theoretical literature in LLM watermarking (Li, Ruan, Wang, Long & Su 2025*b*, Cai et al. 2025, Christ et al. 2024, Li et al. 2024, Li, Ruan, Wang, Long & Su 2025*a*) assumes that either the entire text, or at least a constant proportion of the text, is watermarked.

Assumption 4.1 is mild in the sense it allows for vanishing watermarked patches in the $[0, 1]$ scale. Mathematically, we only require the minimum size of the watermarked patches (as well as the minimum separation) to be grow only polynomially with the number of tokens n. These assumptions are ubiquitous in the analysis of multiple change-point (Fryzlewicz 2014, Cho 2016, Bai et al. 2023, Safikhani & Shojaie 2022, Frick et al. 2014), as well as in the relatively much sparser literature of analysis of multiple epidemic patches (Zhao & Yau 2021, Juodakis & Marsland 2023).

In what follows, we explain the step-by-step rationale behind the algorithm. For clarity, we ignore the niceties of $\lfloor\cdot\rfloor$'s and $\lceil\cdot\rceil$'s. Suppose, for convenience, that $\upsilon = 1/2$.

- **Blocking stage.** For convenience let $b = \sqrt{n}$. In the first stage, we partition the data into $\sqrt{n}$ consecutive blocks of size $\sqrt{n}$. Let the threshold $\mathcal{Q}$ being given as some quantile of the distribution of the maximum of the block-sums of the pivot statistics under the null of no watermarking. Then, among the blocks, we retain only those blocks for which the corresponding realized sum of pivot statistics exceeds $\mathcal{Q}$. Typically, to avoid multiple testing issues, $\mathcal{Q}$ is chosen as the $(1 - \alpha)$-quantile of the *null* (i.e., when there is no watermarking in the entire text) distribution of

the maximum block sum over all $\frac{n}{b}$ blocks.

- **Discarding stage.** If α is too small, we risk selecting many spurious blocks; if α is too large, we lose out on power in the first stage itself, failing to accurately identify even the number of watermarked segments. As a calibration step, we form connected components based on selected blocks, and then remove any of the intervals having length smaller than $c\sqrt{\log n}$, $c > 0$. The intuition is as follows: the blocks corresponding to the un-watermarked region between them should not be selected; else we lose the localization we are aiming for before implementing $\hat{I}$ piece-meal. Moreover, under Assumption 4.1, by definition of $\mathcal{Q}$, $\sqrt{\log n}$ successive un-watermarked blocks will have sums exceeding $\mathcal{Q}$ *only* with vanishing probability. Therefore, any connected interval of selected blocks from the first stage, with length at most $c\sqrt{\log n}$, must necessarily be spurious.
- **Enlargement stage.** The above two steps ensure $\hat{K} = K$ with probability approaching 1. Due to Assumption 4.1, each of the watermarked segments must correspond to exactly one of the remaining connected regions. Moreover, these intervals are almost accurate estimates of the true segments, but for some additional watermarked regions that might have had a non-null intersection with the discarded blocks. However, from the particular discarding procedure, we know that these additional regions must account for a size at most of the order of $\sqrt{n}$. Therefore, it makes sense to enlarge the connected intervals by $c\sqrt{n}$ for some constant $c > 0$, so that now it covers the corresponding true watermarked segments with high probability. These enlarged intervals D_j's remain disjoint with high probability due to Assumption 4.1, and are therefore each amenable to (5) to yield $\hat{I}_j$'s.
- **Estimating $\tilde{d}$.** The crucial component behind $\hat{I}_j$ is $\tilde{d}$, which we estimate now. In fact, we plug in the sample mean of the pivot statistics over $\cup_{j=1}^{\hat{K}} D_j$ as $\tilde{d}$. Since $|D_j \Delta I_j| \ll |I_j|$ with high probability, hence $\tilde{d}$ is essentially equal to $(\sum_j |I_j|)^{-1} \sum_{j\in[K], t\in I_j} (X_t - \mu_0)$, which estimates d with some positive bias. The ρ parameter can be used to calibrate it so that $d > 2\rho\tilde{d}$. Typically

we choose $\rho \in (0.1, 0.5)$. A smaller value of ρ maintains validity of the procedure but sacrifices the detection accuracy. In Appendix §C.3 we provide an ablation study to discuss the choices of both the parameters b and ρ.

- **Reducing computational cost.** We alleviate the increased computational aspect of a naive implementation of (5) by leveraging additional information from the screening stage to reduce the search space. Indeed, due to our blocking and discarding steps, it can be guaranteed with high probability that, for each $j \in [K]$, $D_{j,L}$ is at most $\asymp \sqrt{n}$ distance apart from $I_{j,L}$; similarly $D_{j,R}$ is also at most $\asymp \sqrt{n}$ distance apart from $I_{j,R}$. Therefore, from D_j we can produce search intervals L_j, R_j of lengths $\asymp n^{1/2}$ such that $I_{j,L} \in L_j$ ad $I_{j,R} \in R_j$ with high probability, and restrict the search to $s \in L_j, t \in R_j$. Consequently, now each implementation of this modified (5) (see Figure 2) takes $O((n^{1/2})^2) = O(n)$ amount of computational time, leading to a speed-up while maintaining theoretical validity.

The following result summarizes these insights into a formal consistency guarantee.

Theorem 4.1. *Assume that the null distribution of the pivot statistic is absolutely continuous with respect to the Lebesgue measure. Let the number of watermarked intervals K be bounded, and Assumption 4.1 be granted for the watermarked intervals $I_k, k \in [K]$. Fix $\alpha \in (0, 1)$, and recall the quantities defined in* `WISER` *described in Figure 2. Suppose that $\mathbb{E}_0[|X - \mu_0|^p] < \infty$ for some $p \geq 2$, and let the block length $b = b_n$ satisfy $b_n = O(n^\upsilon)$, and $b_n/n^{1/p} \to \infty$, where $\upsilon > 1/p$ is same as in Assumption 4.1. Moreover, suppose the threshold $\mathcal{Q} = \mathcal{Q}_n$ is selected so that $\mathbb{P}_0(\max_{1 \leq k \leq \lceil n/b \rceil} S_k > \mathcal{Q}) = \alpha$. Finally, assume $d \geq c$ for some constant $c > 0$, and $\sup_{P \in \mathcal{P}} \mathbb{E}_{1,P}[X] < \infty$, and assume there exists $\tau > 0$ such that*

$$\kappa := \inf_{\theta \geq 0} \theta(\mu_0 + \tau d) + \log \sup_P \mathbb{E}_{1,p}[\exp(-\theta X)] < 0. \tag{6}$$

Suppose $\varepsilon > 0$ and $d \geq c$ be given for some constant $c > 0$. Then, under the assumptions of

Theorem 3.1, there exist $M_\varepsilon \in \mathbb{R}_+$*, independent of* n, K*, and* d*, and* $\rho > 0$*, such that* `WISER` *applied with hyper-parameters* b *and* ρ *satisfies*

$$\liminf_{n\to\infty} \mathbb{P}\Big(\hat{K} = K,\ \max_{k\in[K]} |\hat{I}_k \Delta I_k| < M_\varepsilon d^{-1}\Big) \geq 1 - \varepsilon. \tag{7}$$

Remark 4.2 (Effect of Assumption 4.1). The key assumption behind the validity of `WISER` is that if the pivot distribution under null has at least p moments, then the minimum length of the watermarked intervals, as well as the minimum separation between them, should be at least $n^{1/p}$; in that case, `WISER` guarantees consistent segmentation as long as the block-length is small enough. In fact, for most watermarking schemes in use, the pivot distribution under null will have infinitely many moments, enabling us to take as big a p as possible. Thus, our Assumption 4.1 can be understood to be quite mild, much like the logarithmic separation conditions in multiple change-point detection literature (e.g., Assumption 3.3 in Fryzlewicz (2014), Assumption (B2) in Cho (2016), Assumption (H1') in Bai et al. (2023), etc). However, we remark that the aforementioned separation conditions from change-point literature are often proposed under Gaussianity, or under specific dependency structures, none of which hold true for the watermarked interval in our set-up.

Remark 4.3 (Discussion on Condition (6)). We also briefly discuss arguably the only technical condition (6) in Theorem 4.1. This can be construed as a Donsker-Varadhan strengthened version of Assumption 2.2. For an appropriate choice of the score function h and some NTP distribution $P^\star$ depending on $\mathcal{P}$, the Donsker Varadhan representation (Donsker & Varadhan 1983) entails

$$\inf_{\theta\geq 0} \theta\mu_0 + \log \sup_{P\in\mathcal{P}} \mathbb{E}_{1,p}[\exp(-\theta X)] = -D_{\mathrm{KL}}(\mathcal{L}_0(X), \mathcal{L}_{1,P^\star}(X)),$$

where D_{KL} denotes the Kullback-Leibler divergence, $\mathcal{L}_0$ denotes the law of un-watermarked pivot statistic, and $\mathcal{L}_{1,P^\star}$ denotes the law of watermarked pivot statistic when the NTP is $P^\star$. In light of

this, κ lifts the minimum separation between the un-watermarked and watermarked distributions into a gap between the cumulant functions, and can therefore be understood to be mild. Equation (6) establishes a weak uniform control over the behavior of pivot statistics under watermarked segments. This allows us to rigorously bypass the possibly arbitrary and strong dependence across the pivot statistics corresponding to watermarked tokens while deriving Theorem 4.1.

We reiterate that with any choice of $b_n = O(\sqrt{n})$, `WISER` has a run-time only of approximately $O(n)$ ignoring log factors. This, to the best of our knowledge, is among the *least computationally expensive* algorithms available in the literature. In view of its theoretical validity under very general conditions, this makes it a useful tool for practical applications. In general, for consistent segmentation, the block lengths only need to satisfy $n^{1/p} \ll b_n \ll n^{\nu}$, where ν is as in Assumption 4.1, and the pivot statistics has at least p finite moments. In Appendix §C.3, we undertake a detailed ablation study that deals with the practical choice of b_n.

5 Simulation Studies

Building on the theoretical results developed in the preceding sections, we now present a series of simulation experiments designed to stress-test the performance guarantee of the proposed `WISER` method, under deviations from the idealized assumptions. To this end, we consider three distinct simulation scenarios as follows. In §5.1, we aim to uncover the effect of temporal dependence on the performance of the `WISER` algorithm. Moving on, §5.2 examines the role of Assumption 2.2 on the performance of the algorithm. Finally, in §5.3, we evaluate `WISER` on its ability to detect multiple watermarked segments based on simulated text, which then serves as the closest approximation to the additional experiments on real-world scenarios discussed later in §6. For each of these experiments, we keep the vocabulary size fixed at $|\mathcal{W}| = 1000$ unless otherwise specified and perform 5000 replications.

5.1 Effect of temporal dependence on `WISER`

We begin by analyzing how temporal dependence in the underlying next-token prediction (NTP) distributions affects detection accuracy. For each of the 5000 replications, we generate a sequence of NTP distributions $\{P_t\}_{t=1}^T$ according to

$$P_t(w) = \frac{e^{z_t(w)}}{\sum_{w\in\mathcal{W}} e^{z_t(w)}},\ z_t(w) = \sqrt{\phi} z_{t-1}(w) + \sqrt{1-\phi}\sigma_t(w),\ t = 1, 2, \ldots,\ w \in \mathcal{W},$$

where ϕ is the auto-correlation coefficient ranging from 0 to 1, and $\sigma_t(w)$'s are independent, identically distributed logits generated from a spiked-probability distribution described by

$$P_t(w) = (1-\Delta_t)\mathbf{1}_{\{w=w_t^*\}} + \frac{\Delta_t}{|\mathcal{W}|-1}\mathbf{1}_{\{w\neq w_t^*\}}$$

where $w_t^* \sim$ Uniform$(\mathcal{W})$ and $\Delta_t \sim U(10^{-3}, 0.5)$. Note that $\phi = 0$ recovers the case where each NTP is generated as an independent spiked distribution, which is equivalent to the scenario considered by Li, Ruan, Wang, Long & Su (2025*b*).

For each replication, we generate a text of length $n = 500$, watermark the interval $(50, 300)$ under various schemes, and compute the resulting IOU of the intervals detected by `WISER`. The results are summarized in Figure 3. Overall, `WISER` is remarkably stable across a broad range of temporal dependencies, except in two cases: (i) $\phi = 1$, and (ii) red-green watermark scheme with smaller values of ϕ.

In the extreme case $\phi = 1$, all NTP distributions $\{P_t\}$ collapse to the initial spiked distribution P_0. Because of the spiked-nature of P_0, a decoding (or sampling) from this model produces nearly deterministic text consisting almost entirely of the token w_0^*. Therefore, watermarking has virtually no opportunity to influence the output, because deviations from w_0^* occur only through extremely low-probability events. This makes it extremely difficult for any detection algorithm to distinguish

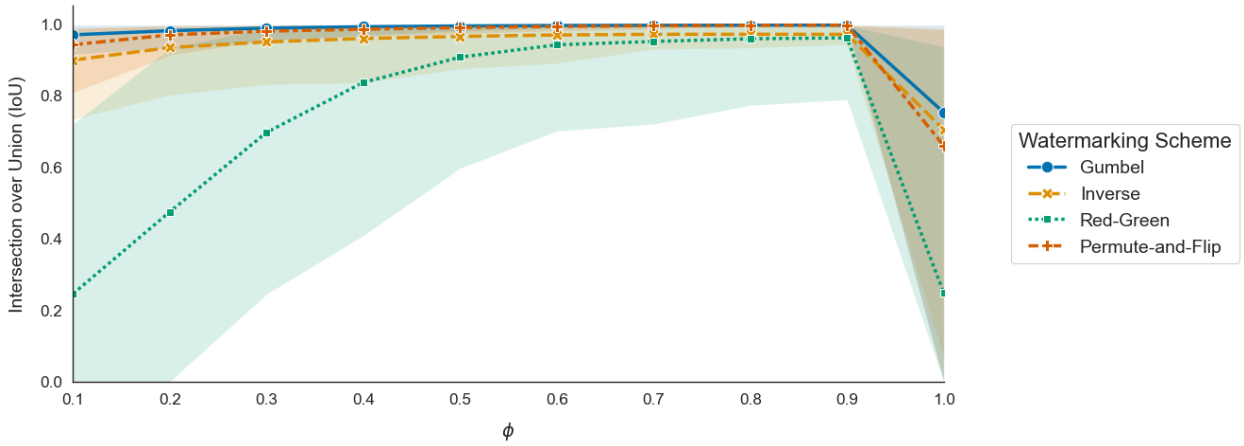

Figure 3: IOU of WISER algorithm for different levels of correlation in the token generation process given in simulation scenario of §5.1

between watermarked and unwatermarked cases. Also, in this degenerate case, because of the deterministic nature of the generated text, the variance $\min\{\text{Var}_0(\epsilon), \sup_P \text{Var}_{1,P}(\epsilon)\}$ reduces to almost zero, violating the assumptions of Theorem 3.1.

On the other extreme, when $\phi = 0$, each spiked distributions are independently generated, producing the maximum probability at different tokens. The red-green scheme perturbs logits by adding a bias only on the tokens from the green list, but this influence is mitigated by the large and rapidly changing spikes. Thus, watermarked and unwatermarked texts become nearly indistinguishable.

5.2 Effect of Assumption 2.2 on WISER

In order to properly characterize the importance of the elevated mean of pivot statistics for watermarked texts, we consider two different simulation experiments. In the first experiment, we watermark a single interval $(0.3n, 0.7n)$ within a text of length n using the red-green scheme. We vary the watermark strength by setting the logit bias $\delta \in 1.5, 2.0, \ldots, 3.5$

In the first one, we apply the red-green watermark on a single interval $(0.3n, 0.7n)$ for a n-length text, with choices of $\delta \in \{1.5, 2, \ldots, 3.5\}$, where δ is the bias added to the logit of the tokens from

the green list. The corresponding IOU values for `WISER` are shown on the top panel of Figure 4.

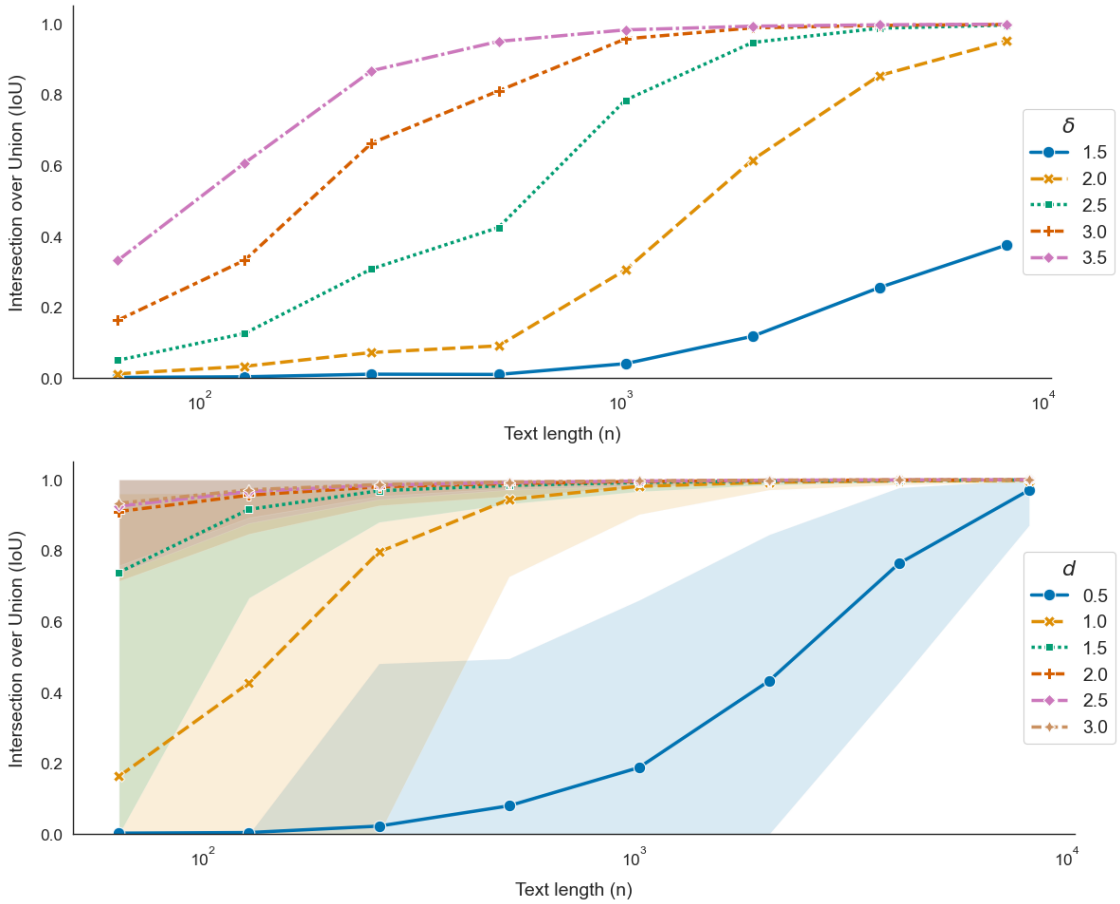

Figure 4: IOU of `WISER` algorithm across different length texts (Top) for different levels of δ in Red-green watermarking; (Bottom) for different levels of d as in Assumption 2.2 for Gumbel watermarking scheme. X-axis is in log-scale for both plots.

On the other hand, such control of the strength of the watermarking is not possible for the Gumbel watermarking scheme. Therefore, as an illustration, we perform another simulation study by directly generating the pivot statistic without any explicit choice of the underlying watermarking scheme. We choose the pivot statistic at the unwatermarked regions as i.i.d. exponentially distributed random variables with mean 1 (note that, this is the distribution of the pivot statistic Y_t corresponding to the Gumbel watermarking scheme for the unwatermarked tokens), and choose the watermarked regions as i.i.d. normal random variables with mean $(1 + d)$ and variance 1. The results for this simulation are illustrated on the bottom panel of Figure 4.

Both the plots present in Figure 4 convey the asymptotic consistency of the algorithm, as well as establish an empirical validity of the error bound given in Theorem 3.1. As d decreases, the problem of watermark detection becomes more difficult, and a larger text length n is required to achieve the same level of IOU.

5.3 Experiments on multiple watermarked segment detection

We next evaluate `WISER` in settings where multiple watermarked intervals appear within the text. Consider text of length n containing three Gumbel-watermarked intervals: $(0.35n - g, 0.45n - g), (0.45n, 0.55n)$ and $(0.55n + g, 0.65n + g)$. Here g denotes the gap between two successive intervals, and is selected as $g = (\tilde{g} \vee 2) \wedge 0.3n$, where $\tilde{g} \in \{n^{0.1}, n^{0.15}, \ldots, n^{0.9}, n^{0.95}\}$.

Theorem 4.1 assures successful detection of the intervals by `WISER` algorithm whenever the gaps satisfy $d(I_k, I_{k-1}) \asymp \sqrt{n}$. Figure 5 displays, as a function of $\log_n(d(I_k, I_{k-1}))$ for different text lengths n, the empirical probability that `WISER` correctly detects three watermarked intervals. As expected, for each of the n considered here, the empirical detection probability rises sharply from 0 to 1 around $\log_n(d(I_k, I_{k-1})) \approx 0.5$, consistent with the theoretical results.

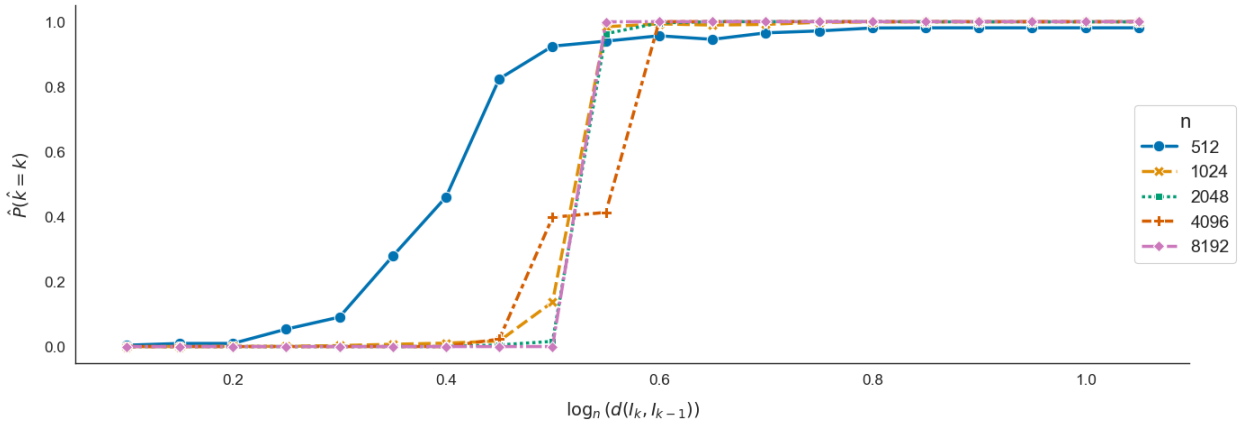

Figure 5: Empirical proportion of repetitions where `WISER` algorithm determines correct number of intervals in simulation scenario §5.3 as a function of the logarithm of gap $\log_n(d(I_k, I_{k-1}))$.

6 Numerical Experiments

While the previous section corroborates the theoretical results with numerical simulations, this section aims to demonstrate the superiority of the proposed `WISER` method over existing state-of-the-art (SOTA) algorithms, when applied in a real-world scenario. In §6.1, we compare its accuracy against competitive methods on a benchmark dataset across multiple watermarking schemes, and in §6.2, we assess its computational efficiency. Due to space constraints, we provide additional numerical experiments in Appendix §C. We encourage the readers to check it out for more practical insights, including, **(i)** a detailed explanation of the benefits of `WISER` over other SOTA algorithms (§C.1.3), **(ii)** experiments quantifying the effect of watermark intensity and length across different algorithms (§C.2), and **(iii)** an ablation study (§C.3) highlighting the stability of our method across tuning parameter choices. The datasets and the large language models were acquired from the open-source Huggingface library. All the relevant reproducible codes and figures can be found in the anonymous Github repository.

6.1 Comparative performance of `WISER`

Within the relatively limited body of literature on the identification of watermarked segments from mixed-source texts, `Aligator` (Zhao, Liao, Wang & Li 2024), `SeedBS-NOT` (Li et al. 2024), and `Waterseeker` (Pan et al. 2025) algorithms have emerged as the leading methods, producing the most accurate results so far. For an extensive comparison, our experimental setup involves completion of randomly selected 200 prompts from the Google C4 news dataset[1]. We include language models spanning a wide range of scales: parameter sizes varying from 125 million to 8 billion, and vocabulary sizes ranging in 32-262 thousands; for watermarking schemes, we consider Gumbel-max trick (Aaronson 2023), Inverse transform (Kuditipudi et al. 2024), Red-green watermark (Kirchenbauer et al. 2023), and Permute-and-Flip watermark (Zhao et al. 2025). In

[1] https://www.tensorflow.org/datasets/catalog/c4

each scenario, the first 50 tokens of a news article have been provided as inputs to the language models, and $n = 500$ output tokens are recorded. Among these 500 output tokens, there are two watermarked segments: 100-200 and 325-400. The specific tuning parameter choices for `WISER` are provided in §C. Table 1 showcases the results for the Gumbel watermarking scheme. It is evident that `WISER` outperforms all the other algorithms across all the metrics for each model. The detailed discussions, including the specific metrics used and additional results and insights, are provided in Appendix §C.1.

Model Name	Vocab Size	Method	IOU	Precision	Recall	F1	RI	MRI
facebook/opt-125m	50272	`WISER`	**0.944**	**1.000**	**0.995**	**0.997**	**0.984**	**0.979**
		`Aligator`	0.734	0.382	0.988	0.551	0.939	0.931
		`Waterseeker`	0.672	1.000	0.802	0.890	0.864	0.850
		`SeedBS-NOT`	0.479	0.730	0.625	0.673	0.844	0.823
google/gemma-3-270m	262144	`WISER`	**0.896**	**0.965**	**0.960**	**0.962**	**0.953**	**0.950**
		`Aligator`	0.506	0.234	0.912	0.373	0.881	0.861
		`Waterseeker`	0.645	0.968	0.775	0.861	0.851	0.836
		`SeedBS-NOT`	0.362	0.610	0.478	0.536	0.753	0.704
facebook/opt-1.3b	50272	`WISER`	**0.934**	**1.000**	**0.995**	**0.997**	**0.981**	**0.974**
		`Aligator`	0.497	0.235	0.920	0.375	0.892	0.871
		`Waterseeker`	0.657	1.000	0.808	0.893	0.860	0.846
		`SeedBS-NOT`	0.360	0.618	0.465	0.531	0.766	0.731
princeton-nlp/Sheared-LLaMA-1.3B	32000	`WISER`	**0.939**	**1.000**	**0.998**	**0.999**	**0.983**	**0.978**
		`Aligator`	0.459	0.236	0.912	0.376	0.886	0.862
		`Waterseeker`	0.659	1.000	0.812	0.897	0.862	0.847
		`SeedBS-NOT`	0.278	0.520	0.388	0.444	0.731	0.699
mistralai/Mistral-7B-v0.1	32000	`WISER`	**0.909**	**1.000**	**0.998**	**0.999**	**0.975**	**0.961**
		`Aligator`	0.292	0.215	0.745	0.334	0.811	0.774
		`Waterseeker`	0.621	1.000	0.765	0.867	0.840	0.824
		`SeedBS-NOT`	0.240	0.442	0.320	0.371	0.657	0.593
meta-llama/Meta-Llama-3-8B	128256	`WISER`	**0.926**	**1.000**	**0.988**	**0.994**	**0.977**	**0.975**
		`Aligator`	0.546	0.367	0.925	0.525	0.911	0.891
		`Waterseeker`	0.570	1.000	0.720	0.837	0.814	0.791
		`SeedBS-NOT`	0.379	0.620	0.515	0.563	0.778	0.741

Table 1: Results for Gumbel Watermarking

6.2 Time Comparison

As established in §4, the proposed `WISER` algorithm achieves a computational complexity of $\approx O(n)$. Figure 6 provides empirical evidence supporting this theoretical claim and, in addition,

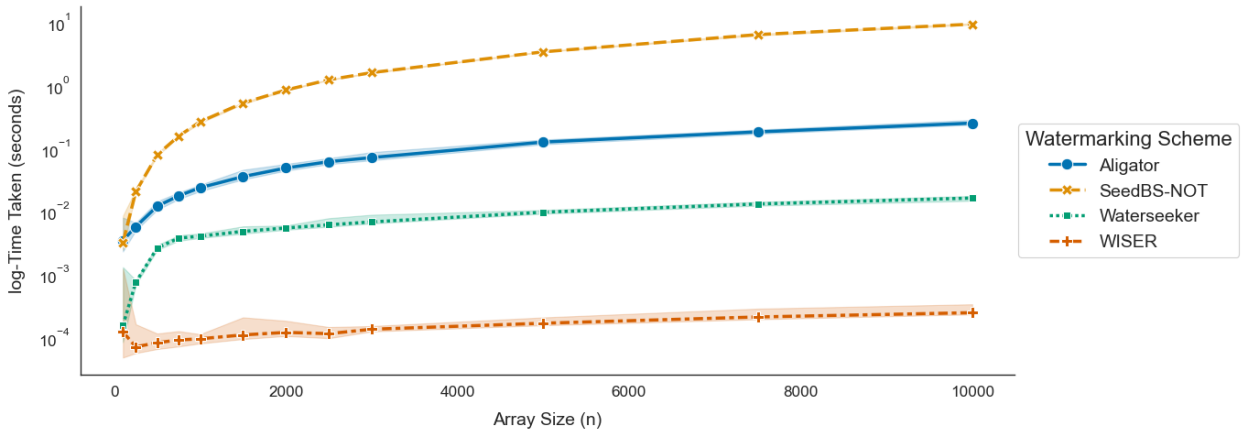

Figure 6: Time complexity (seconds) for various algorithms as a function of completion lengths (n). Y-axis is in log-scale, with 95% confidence interval shown in shades.

compares the runtime behavior of WISER with other state-of-the-art methods. For this experiment, we randomly create an $n/6$-length watermarked segment using the Gumbel-max trick with NTP generated by Google's Gemma-3 model; block size was taken as $\lceil \sqrt{n} \rceil$ and $\rho = 0.1$. The results clearly indicate that WISER consistently outperforms competing approaches in terms of computational efficiency, emerging as the fastest among all methods considered in this study.

7 Conclusion

In this paper, we introduced WISER, a first-of-its-kind algorithm for efficient and theoretically valid segmentation of watermarked intervals in mixed-source texts. By framing watermark localization as an epidemic change-point problem, we bridged a novel connection between classical statistical theory and a modern challenge in generative AI, and also designed a linear time algorithm with provable consistency guarantees, which were further confirmed by our extensive numerical experiments. Beyond the findings of this paper, it is also crucial to theoretically investigate the robustness of the proposed algorithm under human edits (Li, Ruan, Wang, Long & Su 2025*a*); as a roadmap, we have already included some relevant discussion in Appendix §B. Its applicability to multimodal (e.g., audio, image, video) settings (Qiu et al. 2024) also presents opportunities for future research.

8 Data Availability Statement

The prompts used in this paper are publicly available as C4 News dataset with the "tensorflow" package. All the large language models used in Section 6 are available in HuggingFace repository.

References

Aaronson, S. (2023), 'Watermarking of large language models', https://simons.berkeley.edu/talks/scottaaronson-ut-austin-openai-2023-08-17. Talk at the Simons Institute for the Theory of Computing.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. et al. (2023), 'Gpt-4 technical report', *arXiv preprint arXiv:2303.08774* .

Ahmed, A. A. A., Aljabouh, A., Donepudi, P. K. & Choi, M. S. (2021), 'Detecting fake news using machine learning: A systematic literature review', *arXiv preprint arXiv:2102.04458* .

Akiki, C., Pistilli, G., Mieskes, M., Gallé, M., Wolf, T., Ilić, S. & Jernite, Y. (2022), 'Bigscience: A case study in the social construction of a multilingual large language model', *arXiv preprint arXiv:2212.04960* .

Bai, J. (1994), 'Least squares estimation of a shift in linear processes', *J. Time Ser. Anal.* **15**(5), 453–472.

Bai, P., Safikhani, A. & Michailidis, G. (2023), 'Multiple change point detection in reduced rank high dimensional vector autoregressive models', *J. Amer. Statist. Assoc.* **118**(544), 2776–2792.

Bao, G., Zhao, Y., Teng, Z., Yang, L. & Zhang, Y. (2024), Fast-detectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature, *in* 'The Twelfth International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=Bpcgcr8E8Z*

Bartz, D. & Hu, K. (2023), 'Openai, google, others pledge to watermark ai content for safety, white house says', https://www.reuters.com/technology/openai-google-others-pledge-watermark-ai-content-safety-white-house-2023-07-21/. Accessed: 2023-10-03.

Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021), On the dangers of stochastic parrots: Can language models be too big?, *in* 'Proceedings of the 2021 ACM conference on fairness, accountability, and transparency', pp. 610–623.

Biden, J. R. (2023), 'Fact sheet: President biden issues executive order on safe, secure, and trustworthy artificial intelligence', https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/. The White House, October 30, 2023.

Bonnerjee, S., Karmakar, S., Cheng, M. & Wu, W. B. (2025), 'Testing synchronization of change-points for multiple time series', *Preprint* .

Cai, Y., Li, L. & Zhang, L. (2025), 'A statistical hypothesis testing framework for data misappropriation detection in large language models', *arXiv preprint arXiv:2501.02441* .

Cai, Z., Liu, S., Wang, H., Zhong, H. & Li, X. (2024), 'Towards better statistical understanding of watermarking llms', *arXiv preprint arXiv:2403.13027* .

Chen, C. & Shu, K. (2023), 'Can llm-generated misinformation be detected?', *arXiv preprint arXiv:2309.13788* .

Chen, Z., Li, Z. & Zhou, M. (2016), 'Detecting change-points in epidemic models', *Journal of Advanced Statistics* **1**(4), 181.

Cho, H. (2016), 'Change-point detection in panel data via double CUSUM statistic', *Electron. J. Stat.* **10**(2), 2000–2038.

Christ, M., Gunn, S. & Zamir, O. (2024), Undetectable watermarks for language models, *in* S. Agrawal & A. Roth, eds, 'Proceedings of Thirty Seventh Conference on Learning Theory', Vol. 247 of *Proceedings of Machine Learning Research*, PMLR, pp. 1125–1139.

Crothers, E. N., Japkowicz, N. & Viktor, H. L. (2023), 'Machine-generated text: A comprehensive survey of threat models and detection methods', *IEEE Access* **11**, 70977–71002.

Csörgő, M. & Horváth, L. (1997), *Limit theorems in change-point analysis*, Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd., Chichester.

Danskin, J. M. (1967), *The theory of max-min and its application to weapons allocation problems*, Vol. V of *Econometrics and Operations Research*, Springer-Verlag New York, Inc., New York.

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D. & Siemens, G. (2024), 'Impact of ai assistance on student agency', *Computers & Education* **210**, 104967.

Deb, N., Ghosal, P. & Sen, B. (2020), 'Measuring association on topological spaces using kernels and geometric graphs', *arXiv preprint arXiv:2010.01768* .

Donsker, M. D. & Varadhan, S. R. S. (1983), 'Asymptotic evaluation of certain markov process expectations for large time. IV', *Comm. Pure Appl. Math.* **36**(2), 183–212.

Fernandez, P., Chaffin, A., Tit, K., Chappelier, V. & Furon, T. (2023), Three bricks to consolidate watermarks for large language models, *in* '2023 IEEE international workshop on information forensics and security (WIFS)', IEEE, pp. 1–6.

Frick, K., Munk, A. & Sieling, H. (2014), 'Multiscale change point inference', *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76**(3), 495–580. With 32 discussions by 47 authors and a rejoinder by the authors.

Fryzlewicz, P. (2014), 'Wild binary segmentation for multiple change-point detection', *Ann. Statist.* **42**(6), 2243–2281.

Fuk, D. H. & Nagaev, S. V. (1971), 'Probabilistic inequalities for sums of independent random variables', *Teor. Verojatnost. i Primenen.* **16**, 660–675.

Gao, I., Liang, P. & Guestrin, C. (2025), Model equality testing: Which model is this API serving?, *in* 'The Thirteenth International Conference on Learning Representations'.

Gehrmann, S., Strobelt, H. & Rush, A. M. (2019), 'Gltr: Statistical detection and visualization of generated text', *arXiv preprint arXiv:1906.04043* .

Golowich, N. & Moitra, A. (2024), Edit distance robust watermarks via indexing pseudorandom codes, *in* 'The Thirty-eighth Annual Conference on Neural Information Processing Systems'.

Grynbaum, M. M. & Mac, R. (2023), 'The times sues openai and microsoft over ai use of copyrighted work', *The New York Times* **27**(1).

Hájek, J. & Rényi, A. (1955), 'Generalization of an inequality of Kolmogorov', *Acta Math. Acad. Sci. Hungar.* **6**, 281–283.

Hall, P. & Heyde, C. C. (1980), *Martingale limit theory and its application*, Probability and Mathematical Statistics, Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London.

Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J. & Goldstein, T. (2024), 'Spotting llms with binoculars: Zero-shot detection of machine-generated text', *arXiv preprint arXiv:2401.12070* .

Hu, Z., Chen, L., Wu, X., Wu, Y., Zhang, H. & Huang, H. (2024), Unbiased watermark for large language models, *in* 'The Twelfth International Conference on Learning Representations'.

Huang, B., Zhu, B., Zhu, H., Lee, J., Jiao, J. & Jordan, M. (2023), Towards optimal statistical watermarking, *in* 'Socially Responsible Language Modelling Research'.

Hušková, M. & Slabý, A. (1995), 'Testing for an epidemic change in mean', *Commentationes Mathematicae Universitatis Carolinae* **36**(4), 737–747.

Hušková, M. (1995), 'Estimators for epidemic alternatives', *Comment. Math. Univ. Carolin.* **36**(2), 279–291.

Inclán, C. & Tiao, G. C. (1994), 'Use of cumulative sums of squares for retrospective detection of changes of variance', *J. Amer. Statist. Assoc.* **89**(427), 913–923.

Juodakis, J. & Marsland, S. (2023), 'Epidemic changepoint detection in the presence of nuisance changes', *Statistical Papers* **64**(1), 17–39.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I. & Goldstein, T. (2023), A watermark for large language models, *in* A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato & J. Scarlett, eds, 'Proceedings of the 40th International Conference on Machine Learning', Vol. 202 of *Proceedings of Machine Learning Research*, PMLR, pp. 17061–17084.

Kirchenbauer, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., Fernando, K., Saha, A., Goldblum, M. & Goldstein, T. (2024), On the reliability of watermarks for large language models, *in* 'The Twelfth International Conference on Learning Representations'.

Kley, T., Liu, Y. P., Cao, H. & Wu, W. B. (2024), 'Change-point analysis with irregular signals', *Ann. Statist.* **52**(6), 2913–2930.

Kuditipudi, R., Thickstun, J., Hashimoto, T. & Liang, P. (2024), 'Robust distortion-free watermarks for language models', *Transactions on Machine Learning Research* .

Lavergne, T., Urvoy, T. & Yvon, F. (2008), 'Detecting fake content with relative entropy scoring.', *Pan* **8**(27-31), 4.

Lee, J., Le, T., Chen, J. & Lee, D. (2023), Do language models plagiarize?, *in* 'Proceedings of the ACM Web Conference 2023', pp. 3637–3647.

Levin, B. & Kline, J. (1985), 'The cusum test of homogeneity with an application in spontaneous abortion epidemiology', *Statistics in Medicine* **4**(4), 469–488.

Li, R., Allal, L. B., Zi, Y., Muennighoff, N., Kocetkov, D., Mou, C., Marone, M., Akiki, C., Li, J. et al. (2023), 'Starcoder: may the source be with you!', *arXiv preprint arXiv:2305.06161* .

Li, X., Li, G. & Zhang, X. (2024), 'Segmenting watermarked texts from language models', *Advances in Neural Information Processing Systems* **37**, 14634–14665.

Li, X., Liu, X. & Li, G. (2025), 'Adaptive testing for segmenting watermarked texts from language models', *arXiv preprint arXiv:2511.06645* .

Li, X., Ruan, F., Wang, H., Long, Q. & Su, W. J. (2025*a*), 'Robust detection of watermarks for large language models under human edits', *Journal of the Royal Statistical Society Series B: Statistical Methodology* .

Li, X., Ruan, F., Wang, H., Long, Q. & Su, W. J. (2025*b*), 'A statistical framework of watermarks for large language models: pivot, detection efficiency and optimal rules', *Ann. Statist.* **53**(1), 322–351.

Li, X., Wen, G., He, W., Wu, J., Long, Q. & Su, W. J. (2025), 'Optimal estimation of watermark proportions in hybrid ai-human texts', *arXiv preprint arXiv:2506.22343* .

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z. et al. (2024), 'Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews', *arXiv preprint arXiv:2403.07183* .

Liu, Y. & Bu, Y. (2024), Adaptive text watermark for large language models, *in* 'Proceedings of the 41st International Conference on Machine Learning', ICML'24, JMLR.org.

Megías, D., Kuribayashi, M., Rosales, A., Cabaj, K. & Mazurczyk, W. (2022), 'Architecture of a fake news detection system combining digital watermarking, signal processing, and machine learning', *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), 2022, 13 (1): 33-55,* .

Milano, S., McGrane, J. A. & Leonelli, S. (2023), 'Large language models challenge the future of higher education', *Nature Machine Intelligence* **5**(4), 333–334.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D. & Finn, C. (2023), Detectgpt: zero-shot machine-generated text detection using probability curvature, *in* 'Proceedings of the 40th International Conference on Machine Learning', ICML'23, JMLR.org.

Ning, W., Pailden, J. & Gupta, A. (2012), 'Empirical likelihood ratio test for the epidemic change model', *J. Data Sci.* **10**(1), 107–127.

Pan, L., Liu, A., LU, Y., Gao, Z., Di, Y., Huang, S., Wen, L., King, I. & Yu, P. S. (2025),

Waterseeker: Pioneering efficient detection of watermarked segments in large documents, *in* 'AAAI 2025 Workshop on Preventing and Detecting LLM Misinformation (PDLM)'.

Prates, L. d. O. (2021), 'A more efficient algorithm to compute the rand index for change-point problems', *arXiv preprint arXiv:2112.03738* .

Qiu, J., Han, W., Zhao, X., Long, S., Faloutsos, C. & Li, L. (2024), 'Evaluating durability: Benchmark insights into multimodal watermarking', *CoRR* **abs/2406.03728**.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. & Sutskever, I. (2023), Robust speech recognition via large-scale weak supervision, *in* 'Proceedings of the 40th International Conference on Machine Learning', ICML'23, JMLR.org.

Radvand, T., Abdolmaleki, M., Mostagir, M. & Tewari, A. (2025), 'Zero-shot statistical tests for llm-generated text detection using finite sample concentration inequalities', *arXiv preprint arXiv:2501.02406* .

Račkauskas, A. & Suquet, C. (2004), 'Hölder norm test statistics for epidemic change', *J. Statist. Plann. Inference* **126**(2), 495–520.

Račkauskas, A. & Suquet, C. (2006), 'Testing epidemic changes of infinite dimensional parameters', *Stat. Inference Stoch. Process.* **9**(2), 111–134.

Safikhani, A. & Shojaie, A. (2022), 'Joint structural break detection and parameter estimation in high-dimensional nonstationary VAR models', *J. Amer. Statist. Assoc.* **117**(537), 251–264.

Shrestha, Y. R., Von Krogh, G. & Feuerriegel, S. (2023), 'Building open-source ai', *Nature Computational Science* **3**(11), 908–911.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S. et al. (2019), 'Release strategies and the social impacts of language models', *arXiv preprint arXiv:1908.09203* .

Song, Y., Yuan, Z., Zhang, S., Fang, Z., Yu, J. & Liu, F. (2025), Deep kernel relative test for machine-generated text detection, *in* 'The Thirteenth International Conference on Learning Representations'.

Su, J., Zhuo, T. Y., Wang, D. & Nakov, P. (2023), DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text, *in* 'The 2023 Conference on Empirical Methods in Natural Language Processing'.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. et al. (2023), 'Llama: Open and efficient foundation language models', *arXiv preprint arXiv:2302.13971* .

Üstün, A., Aryabumi, V., Yong, Z., Ko, W.-Y., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H.-L., Kayid, A. et al. (2024), Aya model: An instruction finetuned open-access multilingual language model, *in* 'Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', pp. 15894–15939.

Vasilatos, C., Alam, M., Rahwan, T., Zaki, Y. & Maniatakos, M. (2023), 'Howkgpt: Investigating

the detection of chatgpt-generated university student homework through context-aware perplexity analysis', *arXiv preprint arXiv:2305.18226* .

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S. & Wen, Q. (2024), 'Large language models for education: A survey and outlook', *arXiv preprint arXiv:2403.18105* .

Woodcock, C. (2023), 'Ai is tearing wikipedia apart'. Accessed: 2025-09-14.

Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F. et al. (2022), 'Bloom: A 176b-parameter open-access multilingual language model', *arXiv preprint arXiv:2211.05100* .

Wu, Y., Hu, Z., Guo, J., Zhang, H. & Huang, H. (2024), A resilient and accessible distribution-preserving watermark for large language models, *in* 'Proceedings of the 41st International Conference on Machine Learning', ICML'24, JMLR.org.

Xie, Y., Chen, X., Ren, Z. & Su, W. J. (2025), 'Watermark in the classroom: A conformal framework for adaptive ai usage detection', *arXiv preprint arXiv:2507.23113* .

Yao, Q. W. (1993), 'Tests for change-points with epidemic alternatives', *Biometrika* **80**(1), 179–191.

ZeroGPT (2024), 'Trusted gpt-5, chatgpt and ai detector tool'.
URL: *https://www.zerogpt.com/*

Zhao, X., Ananth, P. V., Li, L. & Wang, Y.-X. (2024), Provable robust watermarking for AI-generated text, *in* 'The Twelfth International Conference on Learning Representations'.

Zhao, X., Li, L. & Wang, Y.-X. (2025), Permute-and-flip: An optimally stable and watermarkable decoder for LLMs, *in* 'The Thirteenth International Conference on Learning Representations'.

Zhao, X., Liao, C., Wang, Y.-X. & Li, L. (2024), Efficiently identifying watermarked segments in mixed-source texts, *in* 'Neurips Safe Generative AI Workshop 2024'.

Zhao, Z. & Yau, C. Y. (2021), 'Alternating pruned dynamic programming for multiple epidemic change-point estimation', *Journal of Computational and Graphical Statistics* **30**(3), 808–821.

Zhou, H., Zhu, J., Su, P., Ye, K., Yang, Y., Gavioli-Akilagun, S. A. & Shi, C. (2025), 'Adadetectgpt: Adaptive detection of llm-generated text with statistical guarantees', *arXiv preprint arXiv:2510.01268* .

Zhu, C., Galjaard, J., Chen, P.-Y. & Chen, L. (2024), Duwak: Dual watermarks in large language models, *in* L.-W. Ku, A. Martins & V. Srikumar, eds, 'Findings of the Association for Computational Linguistics: ACL 2024', Association for Computational Linguistics, Bangkok, Thailand, pp. 11416–11436.

Appendix

This appendix is devoted to detailed proofs of our theoretical statements, and additional experimental evidence justifying WISER. §A formally describes the WISER algorithm. In §B, we discuss how Assumption 2.1 can be implemented even in presence of human-edits §C complements the short experimental section in §6 by providing extensive numerical studies concerning the empirical behavior of WISER. Finally, in §D, we provide the detailed mathematical arguments behind WISER.

A WISER algorithm

Algorithm 1: Subroutine Block_Thresholding of WISER

Input: $(X_i)_{i\in[n]}$, block size b, threshold $\mathcal{Q}$

for $k \leftarrow 1$ *to* $\lceil n/b \rceil$ **do**
  $B_k \leftarrow [(k-1)b+1,\ \min(kb,\ n)]$. ;
  $S_k \leftarrow \sum_{l\in B_k} X_l$. ;
$\widetilde{\mathcal{M}} \leftarrow \{\}$. ;
while $k \leq \lceil \sqrt{n} \rceil$ **do**
  if $S_k > \mathcal{Q}$ **then**
    append k to $\widetilde{\mathcal{M}}$;
  $k \leftarrow k+1$. ;
return $\widetilde{\mathcal{M}}$;

Algorithm 2: Subroutine Merging of WISER

Input: $\widetilde{\mathcal{M}} = \{k_1, \ldots, k_m\}$, tuning parameter c

$I \leftarrow \{\}$;
for $j \leftarrow 1$ *to* m **do**
  if j *odd* ***and*** $k_j - 1 \notin K$ ***and*** $k_j + 1 \in K$ **then**
    if $I = \{\}$ ***or*** $k_j - k_{I[end]} > \lceil c\sqrt{\log n} \rceil$ **then**
      append j to I ;
  else if j *even* ***and*** $k_j - 1 \in K$ ***and*** $k_j + 1 \notin K$ **then**
    if $I = \{\}$ ***or*** $k_j - k_{I[end]} > \lceil c\sqrt{\log n} \rceil$ **then**
      append j to I ;
return I;

Algorithm 3: Subroutine `Refined_Local_Search` of `WISER`

Input: $\widetilde{\mathcal{M}} = \{k_1, \ldots, k_m\}$, $I = \{i_1, \ldots, i_{2\hat{K}}\}$, tuning parameters ρ, C

$M \leftarrow \{k_j : j \in I\}$;
Enumerate $M = \{s_1, \ldots, s_{2\hat{K}}\}$;
for $j \leftarrow 1$ *to* $\hat{K}$ **do**
  $D_j \leftarrow [\lfloor (s_{2j-1} - C \log n)b \rfloor \vee 1, \lfloor (s_{2j} + C \log n)b \rfloor \wedge n]$. ;
$\tilde{d} \leftarrow \left(\sum_{j=1}^{\hat{K}} |D_j| \right)^{-1} \sum_{j=1}^{\hat{K}} \sum_{s \in D_j} (X_s - \mu_0)$. ;
for $j \leftarrow 1$ *to* $\hat{K}$ **do**
  $L_j \leftarrow [\lfloor (s_{2j-1} - C \log n)b \rfloor \vee 1, \lfloor (s_{2j-1} + C \log n)b \rfloor]$;
  $R_j \leftarrow (\lfloor (s_{2j} - C \log n)b \rfloor, \lfloor (s_{2j} + C \log n)b \rfloor \wedge n]$. ;
  $\hat{I}_j \leftarrow \arg\min_{s \in L_j, t \in R_j} \sum_{k \in D_j \setminus [s,t]} (X_k - \mu_0 - \rho \tilde{d})$. ;
return *List of estimated watermarked intervals* $\hat{I}_j, j \in [\hat{K}]$.

Algorithm 4: `WISER`

Input: $(X_i)_{i \in [n]}$, block size b, threshold $\mathcal{Q}$, tuning parameters c, C and ρ

$\widetilde{\mathcal{M}} \leftarrow$ `Block_Thresholding`$((X_i)_{i \in [n]}, b, Q)$;
$I \leftarrow$ `Merging`$(\widetilde{\mathcal{M}}, c)$;
$L \leftarrow$ `Refined_Local_Search`$(\widetilde{\mathcal{M}}, I, \rho, C)$;
return *List of estimated watermarked intervals* L;

B Dealing with mixed-source texts

The assumption of knowledge of ζ_t can be too restrictive in most realistic scenarios where human edits are possible. In such cases, one assumes that the pseudo-random numbers ζ_t can also be reconstructed based on the available text and a `Key` with the help of a *Hash function* $\mathcal{A}$:

$$\zeta_t = \mathcal{A}(\omega_{(t-m):(t-1)}, \texttt{Key}). \tag{8}$$

Suppose $\tilde{\omega}_1 \ldots \tilde{\omega}_n$ be a mixed-source text, with segments of un-interrupted watermarked texts punctuated by human-generated texts through substitution, insertion or deletion of LLM generated texts. As a reference, we refer the readers to Procedure 1 of human edits in Li, Ruan, Wang, Long & Su (2025*a*). Note that it is impossible for any verifier to retrieve the exact pseudo-random numbers corresponding to each token in a mixed-source texts. Nevertheless, with the knowledge of

the hash function and Key, one can construct $\tilde{\zeta}_t = \mathcal{A}(\omega_{(t-m):(t-1)}, \texttt{Key})$. Once there is a stretch of un-interrupted watermarked interval with length at least $m \geq 1$, the pseudo-random numbers $\zeta_{t+m}, \zeta_{t+m+1}, \ldots$ can be reliably re-constructed through (8) as the corresponding $\tilde{\zeta}$'s. On the other hand, if ζ_t is not the correct pseudo-random variable associated with ω_t, then either

1. $\tilde{\omega}_t$ is human generated, in which case Working Hypothesis 2.2 of Li, Ruan, Wang, Long & Su (2025*b*) applies to yield ω_t and $\tilde{\zeta}_t$ are independent conditional on P_t;

2. $\tilde{\omega}_t$ is watermarked, which must mean if $\tilde{\omega}_t = \omega_{\tilde{t}}$ in the original watermarked text, then $\omega_{\tilde{t}} = S(P_{\tilde{t}}, \zeta_{\tilde{t}})$ for some true, unknown, pseudo-random number $\zeta_{\tilde{t}}$. In this case we invoke the sensitive nature of the hash function to conclude that ω_t and $\tilde{\zeta}_t$ are independent.

This argument appears in more detail in Section A.1 of Li, Ruan, Wang, Long & Su (2025*a*). In conclusion, the verifier can always obtain access to a sequence $\zeta_{m:n}$ corresponding to a given text $\omega_{1:n}$ such that (i) if $\omega_{(t-m):t}$ is NOT watermarked then ω_t and ζ_t are independent conditional on P_t; (ii), otherwise, ζ_t and ω_t may be intricately dependent on each other. This latter observation is crucial to our subsequent analysis and proposals, for it allows us to construct valid pivotal statistics. In light of the above discussion, we can be excused in making the Assumption 2.1.

Assumption 2.1 can be seen through the lens of constructing the $\tilde{\zeta}_t$ with $m = 1$. We make this slightly simplistic assumption to avoid the un-necessary measure theoretical niceties which might potentially cloud the novelty of our approach. Even with this assumption, proposing a computationally efficient algorithm and establishing its theoretical validity in a setting with multiple watermarked intervals, is an arguably non-trivial task in itself, and to the best of our knowledge, our paper is the first one to deal with this problem with full mathematical rigor. Finally, for a general mixed-source text, we remark that the WISER algorithm can be trivially extended to the setting with general $m \in \mathbb{N}$ by padding an interval of length m to the left of the watermarked segments located by WISER. In the following, we include a further discussion on mixed-source texts along with data misappropriation.

B.1 Further discussion on Assumption 2.2

A general way to deal with mixed-source texts has been proposed in Cai et al. (2025). Therein, the authors allow for modifications to the watermarked tokens by allowing that $d_{\mathrm{TV}}(\omega_t, S(P_t, \zeta_t)) > 0$, as long as the distance is not too large. In light of this, Assumption 2.2 can be further appended by the following:

(B): Let ω be the token generated by modifying $\omega' := S(P, \zeta)$, and let $h(Y), h(Y')$ be accordingly defined via ω, ω'. Then it holds that

$$\inf_{P \in \mathcal{P}} |\mathbb{E}_1[h(Y)] - \mathbb{E}_{1,P}[h(Y')]| \leq d\tau, \ \tau \in (0, 1),$$

where $\mathcal{P}$ and d are same as in Assumption 2.2.

Assumption **(B)** along with Assumption 2.2 ensure that even under potential modification, some degree of separation (characterized by $d(1-\tau)$) remain between the means of the distribution of pivot statistics under null and under watermarking. It is conceivable that our algorithm is consistent even for this case, and the theoretical guarantees should follow more or less similarly; however, to keep the discussion focused, and to convey the key takeaways unhindered, we restrict ourselves to Assumption 2.2.

C Extended numerical experiments

In this section we provide additional numerical experiments complementing those in §6. In §C.1, we compare the accuracy of `WISER` with other competitive methods in the literature, on various benchmark datasets on myriad standard watermarking schemes. Moving on to §C.2, we investigate the effect of watermark intensity as well as the watermarked length on the performance of the algorithms . Finally, in §C.3, we provide some ablation studies corresponding to the hyper-

parameters in WISER.

C.1 Comparative performance of WISER

From §6.1, recall the experimental set-up, the SOTA benchmark algorithms as well as the considered watermarking schemes. For each of the experiments, we implement WISER with block size equal to $b = 65$, $\rho = 0.5$, $\alpha = 0.05$ and $\gamma = 0.1$. Before we provide detailed comparison studies, we elaborate on the performance metrics used.

C.1.1 Performance metric

To ensure consistency with the prior works, we primarily treat the intersection-over-union (IOU) as a performance measure. Let, $\boldsymbol{I} := (I_1, \ldots, I_K)$ denotes the true watermarked intervals and $\widehat{\boldsymbol{I}} := (\widehat{I}_1, \ldots, \widehat{I}_{\hat{K}})$ be the estimated watermarked segments. Then, the intersection-over-union metric is given by

$$\text{IOU}(\boldsymbol{I}, \widehat{\boldsymbol{I}}) = \frac{|(\cup_{i=1}^{K} I_i) \cap (\cup_{j=1}^{\hat{K}} \hat{I}_i)|}{|(\cup_{i=1}^{K} I_i) \cup (\cup_{j=1}^{\hat{K}} \hat{I}_i)|}.$$

Owing to Theorem 3.1, it is obvious that the IOU measure is expected to be close to 1 for the WISER method. Following the definition of Pan et al. (2025), we also compute the precision, recall and F1-score based on whether any of the estimated intervals have a nonempty intersection with any of the true intervals, i.e.,

$$\text{Precision } = \frac{|\{i : 1 \leq i \leq \hat{K}, \hat{I}_i \cap (\cup_{j=1}^{K} I_j) \neq \phi\}|}{\hat{K}}, \text{ Recall} = \frac{|\{i : 1 \leq i \leq \hat{K}, \hat{I}_i \cap (\cup_{j=1}^{K} I_j) \neq \phi\}|}{K}.$$

Rand Index and asymmetry of the watermark segmentation.

In addition to these metrics, the Rand Index (RI) is also usually used to measure coherence between the estimated and true watermarked segments, using the algorithm illustrated in Prates (2021). For

the standard definition of Rand Index, see Equation (2) of Prates (2021). However, the Rand Index may depict a wrong picture of the performance of a watermarked segment identification algorithm. Although watermark segmentation closely resembles epidemic change-point detection, a crucial difference arises in algorithm evaluation. Before proceeding, we briefly deliberate on these issues.

Standard change-point problems are symmetric; under model (3), the edge cases $p = 1, q = n$ and $p = q$ are equivalent. On the other hand, watermarking problems exhibit asymmetry; the edge cases (i) "the entire sequence is un-watermarked" and (ii) "the entire sequence is watermarked", differ due to irregular means of the pivot statistics under watermarking. Rand Index (RI) - despite being used in watermark segmentation (Li et al. 2024, Pan et al. 2025) - fails to capture this distinction.

As an illustration, consider the situation where most of the tokens (say 90%) are watermarked, while the watermark detection algorithm fails to detect any watermarked segment. While the performance of such an algorithm should reflect poorly, the standard Rand Index fails to capture this due to the exchangeability of the watermarked segment and the non-watermarked segment: any pair of indices (i, j) that is truly watermarked trivially is also part of the estimated non-watermarked segment and considered as a concordant pair.

To circumvent these limitations described there, we consider a Modified Rand Index (MRI) given as

$$\begin{aligned}\mathrm{MRI}(\boldsymbol{I}, \widehat{\boldsymbol{I}}) := \mathrm{RI}(\boldsymbol{I}, \widehat{\boldsymbol{I}}) \\ - \frac{\sum_{i \neq j} \left(\sum_{k=1}^{K} \mathbf{1}\{\{i,j\} \subseteq I_k \cap (\cup_{l=1}^{\hat{K}} \hat{I}_l)^c\} + \sum_{l=1}^{\hat{K}} \mathbf{1}\{\{i,j\} \subseteq \hat{I}_l \cap (\cup_{k=1}^{K} I_k)^c\} \right)}{\binom{n}{2}},\end{aligned}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, and n is the number of tokens. The MRI simply adjusts the RI by restricting its exchangeability only within each of the watermarked or non-watermarked intervals, but not in between. Intuitively, the MRI removes the specific pairs of indices (i, j) from the calculation of RI for which both the indices i and j lie either in a true watermarked interval but

are estimated to be in the non-watermarked region, or are estimated to be in a watermarked interval but actually lie in a non-watermarked region.

C.1.2 Experimental results and explanation

The comparison results are summarized in Tables 1, 2-4, corresponding to each of the watermarking schemes considered. Across all watermarking settings, `WISER` consistently delivers the strongest performance across every model and metric. In the Gumbel case, it achieves near-perfect results with IOU scores above 0.90, precision of 1.0, and recall above 0.98 across both small and large models. Competing methods like `Aligator` and `SeedBS-NOT` often fail to balance recall and precision, either collapsing to very low precision (`Aligator`) or producing weaker recall (`SeedBS-NOT`), while `Waterseeker` attains moderate balance but still lags well behind `WISER`.

The trend is even more pronounced in the cases of Inverse and Red-Green setups, where the pivot statistics remain uniformly bounded. In these cases, `Aligator` fail to detect any watermarked intervals, while both `SeedBS-NOT` and `Waterseeker` suffer a significant decline in performance. In contrast, `WISER` maintains F1-scores in the range of 0.95 - 0.99 with stable IOU values across model sizes, showing robustness to different architectures and vocabulary sizes. `Waterseeker` provides the next best alternative, but with noticeable drops in IOU and F1, especially for larger models. These findings clearly demonstrate that `WISER` not only generalises across watermarking schemes but also offers substantial gains in both detection accuracy and reliability, marking a clear benefit over existing baselines.

C.1.3 Why `WISER` outperforms other methods

The enhanced performance of `WISER` does not come out-of-the-blue, rather we argue that it is a byproduct of our unique, epidemic change-point perspective that marries theoretical validity with practical insights. While these methods—`SeedBS-NOT`, `Aligator`, and `Waterseeker`—

Model Name	Vocab Size	Method	IOU	Precision	Recall	F1	RI	MRI
facebook/opt-125m	50272	WISER	**0.906**	**0.995**	**0.980**	**0.987**	**0.968**	**0.931**
		Aligator	0.000	0.000	0.000	0.000	0.203	0.141
		Waterseeker	0.558	0.988	0.710	0.826	0.804	0.783
		SeedBS-NOT	0.178	0.282	0.228	0.252	0.529	0.428
google/gemma-3-270m	262144	WISER	**0.874**	**0.965**	**0.958**	**0.961**	**0.945**	**0.934**
		Aligator	0.000	0.000	0.000	0.000	0.203	0.141
		Waterseeker	0.547	0.983	0.695	0.814	0.797	0.775
		SeedBS-NOT	0.221	0.316	0.272	0.293	0.575	0.544
facebook/opt-1.3b	50272	WISER	**0.846**	**0.980**	**0.928**	**0.953**	**0.934**	**0.904**
		Aligator	0.000	0.000	0.000	0.000	0.203	0.141
		Waterseeker	0.555	0.985	0.698	0.817	0.802	0.781
		SeedBS-NOT	0.189	0.322	0.250	0.282	0.526	0.437
princeton-nlp/Sheared-LLaMA-1.3B	32000	WISER	**0.656**	**0.990**	**0.962**	**0.976**	**0.871**	**0.850**
		Aligator	0.000	0.000	0.000	0.000	0.203	0.141
		Waterseeker	0.582	0.992	0.750	0.854	0.826	0.807
		SeedBS-NOT	0.181	0.286	0.235	0.258	0.541	0.515
mistralai/Mistral-7B-v0.1	32000	WISER	**0.718**	**0.935**	**0.822**	**0.875**	**0.859**	**0.847**
		Aligator	0.000	0.000	0.000	0.000	0.203	0.141
		Waterseeker	0.590	0.996	0.760	0.862	0.830	0.813
		SeedBS-NOT	0.154	0.265	0.192	0.223	0.502	0.489
meta-llama/Meta-Llama-3-8B	128256	WISER	**0.878**	**0.995**	**0.965**	**0.980**	**0.955**	**0.913**
		Aligator	0.000	0.005	0.002	0.003	0.205	0.144
		Waterseeker	0.510	0.980	0.652	0.783	0.774	0.748
		SeedBS-NOT	0.143	0.242	0.185	0.210	0.511	0.480

Table 2: Results for Inverse Watermarking

each contribute useful perspectives, they also exhibit important limitations that the generality of our method usually overcomes.

Limitations of `SeedBS-NOT`: The limitations of `SeedBS-NOT` primarily arise from its reliance on a permutation-based change-point detection framework, which is inherently computationally expensive. Moreover, nowhere they restrict their attention to the specific scenario of watermarked segments, which consigns the change-points to occur in pairs, corresponding to the start and end of a watermarked segment. This is automatically alleviated by `WISER` through its adoption of a natural epidemic change-point formulation. This structural assumption substantially reduces the search space, yielding both computational efficiency and improved statistical stability. Additionally, `SeedBS-NOT` works with the sequence of p-values that are computed from a single observation of the pivot statistic at that location. Due to the complicated nature of the dependence between these

Model Name	Vocab Size	Method	IOU	Precision	Recall	F1	RI	MRI
facebook/opt-125m	50272	WISER	**0.853**	**1.000**	**0.975**	**0.987**	**0.914**	**0.903**
		Aligator	0.000	0.000	0.000	0.000	0.259	0.209
		Waterseeker	0.730	0.998	0.815	0.897	0.889	0.882
		SeedBS-NOT	0.570	0.665	0.615	0.639	0.897	0.870
google/gemma-3-270m	262144	WISER	**0.838**	**0.973**	**0.970**	**0.972**	**0.908**	**0.896**
		Aligator	0.000	0.000	0.000	0.000	0.203	0.141
		Waterseeker	0.643	0.982	0.820	0.894	0.864	0.850
		SeedBS-NOT	0.600	0.749	0.738	0.743	0.900	0.872
facebook/opt-1.3b	50272	WISER	**0.846**	**0.993**	**0.990**	**0.992**	**0.923**	**0.913**
		Aligator	0.000	0.000	0.000	0.000	0.203	0.141
		Waterseeker	0.623	0.990	0.815	0.894	0.851	0.836
		SeedBS-NOT	0.597	0.764	0.735	0.749	0.901	0.874
princeton-nlp/Sheared-LLaMA-1.3B	32000	WISER	**0.850**	**1.000**	**0.990**	**0.995**	**0.919**	**0.908**
		Aligator	0.000	0.000	0.000	0.000	0.203	0.141
		Waterseeker	0.619	0.995	0.810	0.893	0.851	0.836
		SeedBS-NOT	0.570	0.775	0.738	0.756	0.898	0.860
mistralai/Mistral-7B-v0.1	32000	WISER	**0.814**	**0.995**	**0.955**	**0.975**	**0.909**	**0.898**
		Aligator	0.000	0.000	0.000	0.000	0.203	0.141
		Waterseeker	0.559	0.993	0.742	0.850	0.818	0.799
		SeedBS-NOT	0.507	0.718	0.672	0.695	0.877	0.843
meta-llama/Meta-Llama-3-8B	128256	WISER	**0.864**	**1.000**	**0.995**	**0.997**	**0.929**	**0.919**
		Aligator	0.000	0.000	0.000	0.000	0.203	0.141
		Waterseeker	0.647	1.000	0.838	0.912	0.866	0.851
		SeedBS-NOT	0.590	0.778	0.770	0.774	0.919	0.883

Table 3: Results for Red-Green Watermarking

p-values, they are difficult to combine to increase the statistical power. Our approach circumvents this by aggregating the pivot statistic at the block level (Step 7 in Figure 2), enhancing the effective sample size and increasing the power of the detection.

Limitations of `Aligator`: The `Aligator` algorithm frames the task as a reinforcement learning problem, producing a smoothed estimate of the underlying generative process and subsequently applying token-level hypothesis tests with a p-value threshold. While this strategy can capture localized deviations, it often results in a large number of short and fragmented detections, many of whom might be spurious due to possible multiple testing. Consequently, the method tends to produce many disjoint intervals, which severely diminishes its precision. By contrast, the discarding stage of `WISER` enforces structural coherence at the segment level, before returning fine-grained estimate through applying (5). This ensures that localized intervals correspond more closely to

Model Name	Vocab Size	Method	IOU	Precision	Recall	F1	RI	MRI
facebook/opt-125m	50272	WISER	**0.925**	**0.998**	**0.998**	**0.998**	**0.980**	**0.979**
		Aligator	0.665	0.345	0.978	0.510	0.935	0.927
		Waterseeker	0.712	1.000	0.905	0.950	0.891	0.884
		SeedBS-NOT	0.469	0.725	0.560	0.632	0.867	0.817
google/gemma-3-270m	262144	WISER	**0.935**	**1.000**	**1.000**	**1.000**	**0.982**	**0.973**
		Aligator	0.558	0.252	0.952	0.399	0.906	0.889
		Waterseeker	0.614	1.000	0.782	0.878	0.841	0.824
		SeedBS-NOT	0.334	0.610	0.440	0.511	0.766	0.686
facebook/opt-1.3b	50272	WISER	**0.904**	**1.000**	**0.990**	**0.995**	**0.972**	**0.969**
		Aligator	0.446	0.216	0.928	0.350	0.887	0.863
		Waterseeker	0.677	1.000	0.840	0.913	0.873	0.861
		SeedBS-NOT	0.350	0.573	0.430	0.491	0.753	0.717
princeton-nlp/Sheared-LLaMA-1.3B	32000	WISER	**0.919**	**1.000**	**1.000**	**1.000**	**0.979**	**0.977**
		Aligator	0.397	0.202	0.870	0.328	0.870	0.842
		Waterseeker	0.653	1.000	0.778	0.875	0.851	0.837
		SeedBS-NOT	0.264	0.486	0.350	0.407	0.688	0.666
mistralai/Mistral-7B-v0.1	32000	WISER	**0.896**	**1.000**	**0.998**	**0.999**	**0.973**	**0.972**
		Aligator	0.215	0.164	0.672	0.263	0.817	0.774
		Waterseeker	0.646	1.000	0.795	0.886	0.853	0.838
		SeedBS-NOT	0.238	0.468	0.315	0.376	0.650	0.575
meta-llama/Meta-Llama-3-8B	128256	WISER	**0.908**	**1.000**	**0.998**	**0.999**	**0.976**	**0.976**
		Aligator	0.551	0.351	0.950	0.513	0.911	0.891
		Waterseeker	0.535	1.000	0.712	0.832	0.799	0.773
		SeedBS-NOT	0.413	0.658	0.545	0.596	0.775	0.730

Table 4: Results for Permute and Flip Watermarking

contiguous watermark insertions.

Limitations of `Waterseeker`: The `Waterseeker` algorithm may seem structurally similar to the proposed `WISER` method, in that it also employs a two-stage detection framework. However, `Waterseeker` considers a sliding window-based testing mechanism in its first stage, which has a crucial limitation. Consider a very realistic scenario when one of the pivot statistics corresponding to an un-watermarked token is high simply due to random chance. In `Waterseeker`, this will push the score up for W consecutive windows, usually resulting in a false positive in the first stage. On the other hand, for `WISER` such anomalous pivot statistics will affect only one block, which, being usually part of a connected interval with small length, can potentially be discarded with a very high probability in our *discarding stage*. For larger model, this scenario is extremely likely, making this reduction in precision much more pronounced (see Models google/gemma-3-270m and

meta-Ilama/Meta-Llama-3-8B in Tables 1, 2 - 4). Moreover, Pan et al. (2025) provide only limited theoretical validation of their approach, making the optimal tuning of hyper-parameters difficult to justify. This lack of statistical guarantees limits its reliability across watermarking schemes and model sizes, in contrast to the rigorous and general guarantees underlying `WISER`.

C.2 Effect of Watermark Intensity

Type	Method	IOU	F1	RI	MRI
Strong but short	`WISER`	0.794	0.984	0.933	0.925
	`SeedBS-NOT`	0.639	0.785	0.919	0.900
	`Waterseeker`	0.878	0.997	0.969	0.967
Weak but long	`WISER`	0.745	0.779	0.628	0.551
	`SeedBS-NOT`	0.172	0.321	0.675	0.187
	`Waterseeker`	0.268	0.847	0.519	0.172

Table 5: Effect on watermarking signal strength

Following the experimental design of Pan et al. (2025), we evaluate the comparative performance of the proposed `WISER` algorithm under varying levels of watermark intensity. As a demonstration, we choose Google's Gemma-3 series model (270 million) to generate a completion of 500 tokens for each input prompt. The watermark strength is modulated through the bias parameter δ of the Red-Green watermarking scheme (Kirchenbauer et al. 2023), while another parameter m specifies the length of the watermarked region by applying the decoding strategy to the middlemost m tokens within the 500-token output.

In the "strong but short" configuration ($\delta = 2.0, m = 100$), as shown in Table 5, all methods perform well, achieving a Rand Index exceeding 0.9. Although `WISER` is not the best-performing method in this particular case, it remains competitive with `Waterseeker`, which achieves the highest score. By contrast, in the "weak but long" configuration ($\delta = 1.0, m = 400$), only `WISER` maintains robust performance. While `SeedBS-NOT` appears to achieve a higher Rand Index, this

outcome is primarily attributed to the issues described in §C.1.1. The Modified Rand Index (MRI) offers a more reliable assessment, highlighting the superiority of WISER in this setting.

C.3 Ablation studies

We also perform an ablation study to understand the effectiveness of the hyper-parameters (e.g. - block size and ρ) of WISER. Our results are arguably quite stable across wide choices of the tuning parameters; nevertheless we provide more informed choices along with additional insights.

For this study, we consider a single watermarked segment from token index 100 to 200, fix $\rho = 0.25$ and vary the tuning parameter b of the WISER algorithm. As one would have hoped, increasing the block size too much decreases the performance, as the smaller watermarked segments gets subsumed in the noise of unwatermarked segments when block sizes are too large. On the other hand, decreasing the block size would reduce the statistical power of the detection algorithm in the first stage itself. Therefore, one requires a judicious choice of the block size to optimally balance these two aspects, which is empirically observed through the upper plot of Figure 7. Based on empirical evidence, we recommend the choice $b \in (\lceil\sqrt{n}\rceil, 3\lceil\sqrt{n}\rceil)$, which works quite well in various settings that we have experimented with, while being also theoretically supported.

A similar conclusion also holds for the choice of ρ, for which we fix the block size as $b = 25$ and vary the tuning parameter ρ. As the choice of d in Assumption 2.2 is exogeneously determined based on the language model and watermarking scheme, a large value of ρ would imply a smaller $\tilde{d}$ and by virtue of Theorem 3.1 would imply a larger error. The lower plot of Figure 7 demonstrates this empirically. However, any value of ρ between 0.1 and 0.5 provides reasonable and relatively stable estimates.

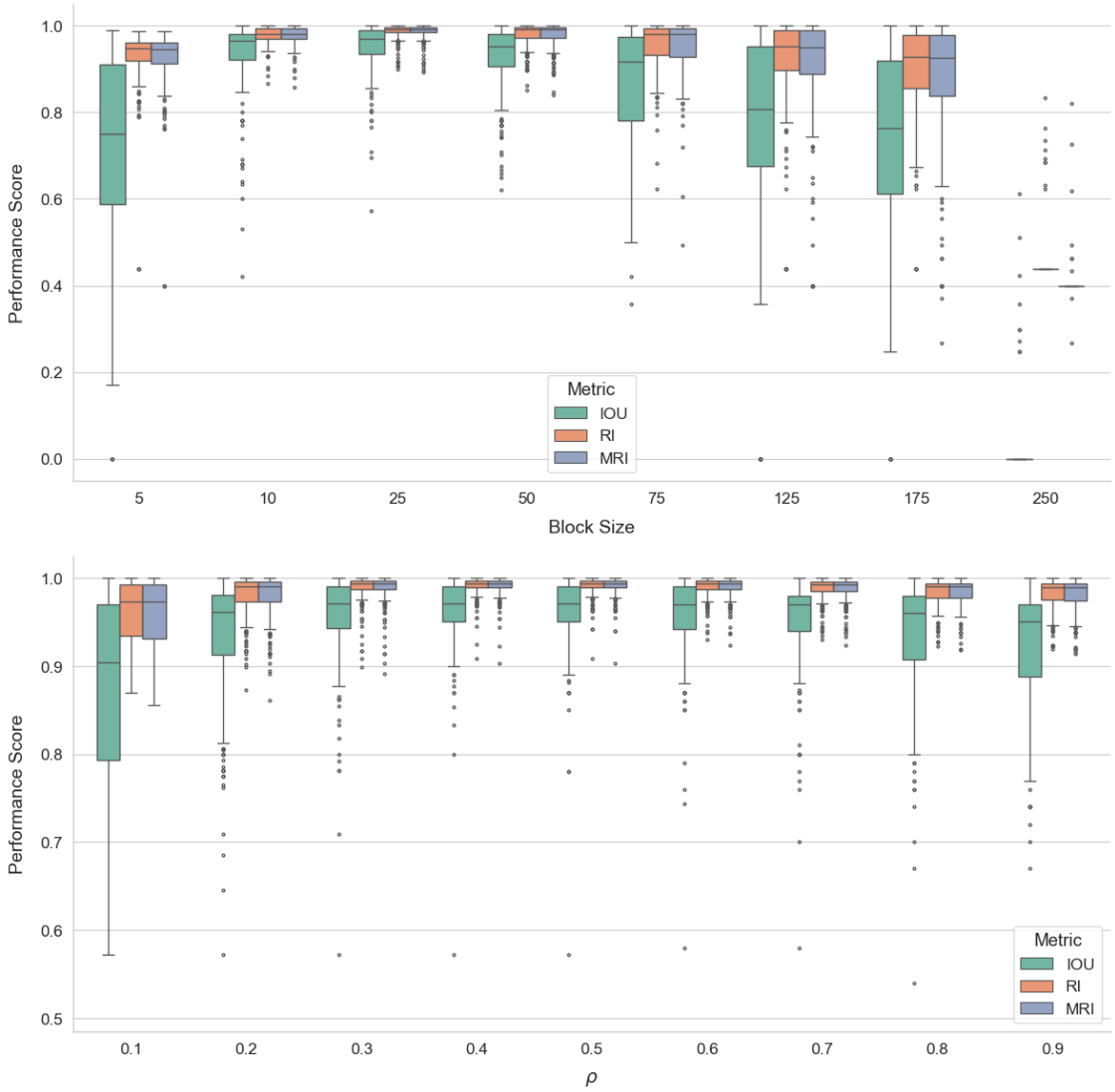

Figure 7: Effect on performance metrics (IOU and Rand Index) due to modification of the hyper-parameters of the `WISER` algorithm, namely block size (Top) and ρ (Bottom).

D Proof of Theoretical Results

In this section, we collect the proofs of theoretical results in the §3. Before we proceed further, we establish some notations. In the following, we write $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for some constant $C > 0$, and $a_n \asymp b_n$ if $C_1 b_n \leq a_n \leq C_2 b_n$ for some constants $C_1, C_2 > 0$. Often we denote $a_n \lesssim b_n$ by $a_n = O(b_n)$. Additionally, if $a_n/b_n \to 0$, we write $a_n = o(b_n)$. For a function $f : \mathbb{R}^n \otimes \mathbb{R}^m \to \mathbb{R}$, let $f^{(1)}(\theta, w) = \frac{\partial}{\partial\theta} f(\theta, w)$, $\theta \in \mathbb{R}^n, w \in \mathbb{R}^m$, $n, m \geq 1$, be the partial derivative function with

respect to θ.

D.1 Proof of Theorem 3.1

In the following, we first state and prove a more generalized version of Theorem 3.1.

Theorem D.1. *Let $\{X_t\}_{t=1}^n := \{h(Y_t)\}_{t=1}^n$ be the pivot statistics based on the given input text, and assume that $I_0 \subset \{1, \ldots, n\}$ be the watermarked interval. Grant Assumption 2.2. Let us also denote*

$$\varepsilon_t = \begin{cases} X_t - \mu_0, t \notin I_0, \\ X_t - \mu_t,\ \mu_t := \mathbb{E}_{1,P_t}[X_t], t \in I_0. \end{cases}$$

Suppose the class of distributions $\mathcal{P}$ is closed and compact, and there exists $\eta > 0$ such that $\sup_{P\in\mathcal{P}} \mathbb{E}_{1,P}[\exp(\eta|\varepsilon|)] < \infty$. Moreover, assume that $\min\{\mathrm{Var}_0(\varepsilon), \sup_P \mathrm{Var}_{1,P}(\varepsilon)\} > 0$. Then it holds that

$$|\hat{I}\Delta I_0| = O_{\mathbb{P}}\Big((\sup_{\theta\geq 0}\{\theta\rho\tilde{d} - \Psi(\theta)\})^{-1}\Big),$$

where Δ denotes the symmetric difference operator, $O_{\mathbb{P}}$ hides constants independent of n and $\tilde{d}$, and

$$\Psi(\theta) = \log \mathbb{E}_0[\exp(\theta\varepsilon)] + 2^{-1}\log\sup_P \mathbb{E}_{1,P}[\exp(2\theta\varepsilon)] + 2^{-1}\log\sup_P \mathbb{E}_{1,P}[\exp(-2\theta\varepsilon)].$$

Theorem D.1 is proved by showing that the probability $\mathbb{P}(|\hat{I}\Delta I_0| > M)$ is small for all sufficiently large M. This probability is controlled by considering the objective function $V_I = S_{I^c} - (\mu_0 + \rho\tilde{d})|I^c|$, where $S_I = \sum_{k\in I} X_k$ and $S_{I^c} = \sum_{k=1}^n X_k - S_I$, and noting that, by construction of $\hat{I}$, $\mathbb{P}(|\hat{I}\Delta I_0| > M) \leq \mathbb{P}(\inf_{I:|I\Delta I_0|>M} V_I - V_{I_0} \leq 0)$. Usually, in change-point literature, one controls terms such as $\inf_{I:|I\Delta I_0|>M} V_I - V_{I_0}$ through Hàjek-Rényi type inequality Hájek & Rényi (1955); see Bai (1994), Bonnerjee et al. (2025). Such inequalities are usually derived by dividing the domain, on which infimum is taken, into smaller intervals, and applying Doob's inequality or Rosenthal's inequality

piece-meal. However, the main bottleneck in a this particular setting is the potentially strong dependence between the pivot statistics in watermarked patches. We develop novel arguments that exploits $\sup_{P\in\mathcal{P}} \mathbb{E}_{1,P}[\exp(\eta|\varepsilon|)] < \infty$ to provide an extended version of the Hajek-Renyi theory through the lens of cumulant generating function. The proof is provided below.

Proof of Theorem D.1. For a candidate watermarked interval I, let $A_1(I) = I \cap I_0^c$, $A_2(I) = I \cap I_0$, $A_3(I) = I^c \cap I_0$, $A_4(I) = (I \cup I_0)^c$, and correspondingly $x_i(I) = |A_i(I)|$, $i = 1(1)4$. Subsequently, we omit the argument I when it is clear from the context. Note that $|I_0| = x_2 + x_3$, $|I| = x_1 + x_2$, and $|\hat{I}\Delta I_0| = x_1 + x_3$. Note that, by definition of $\hat{I}$ it follows that $V_{\hat{I}} \leq V_{I_0}$. Finally, denote $S_i = \sum_{k\in A_i} X_k$, and $S_i^\varepsilon = \sum_{k\in A_i} \varepsilon_k$. With these notations established, we proceed through the following series of implications.

$$
\begin{aligned}
V_I - V_{I_0} &= (S_{I^c} - S_{I_0^c}) - (|I^c| - |I_0^c|)(\mu_0 + \rho\tilde{d}) \\
&= S_3 - S_1 - (x_3 - x_1)(\mu_0 + \rho\tilde{d}) \\
&= (S_3^\varepsilon + \sum_{t\in A_3} \mu_t) - (S_1^\varepsilon + x_1\mu_0) - (x_3 - x_1)(\mu_0 + \rho\tilde{d}) \\
&= S_3^\varepsilon - S_1^\varepsilon + \sum_{t\in A_3} (\mu_t - \mu_0) + (x_1 - x_3)\rho\tilde{d} \\
&\geq S_3^\varepsilon - S_1^\varepsilon + x_3(d - \rho\tilde{d}) + x_1\rho\tilde{d} \qquad (9) \\
&\geq S_3^\varepsilon - S_1^\varepsilon + (x_1 + x_3)\rho\tilde{d}, \qquad (10)
\end{aligned}
$$

where (9) follows from Assumption 2.2 and (10) uses $d \geq 2\rho\tilde{d}$. For some $M > 0$, let $D_M := \{I : |I\Delta I_0| > M\}$. Let $I_0 = [L, R]$. Note that, a candidate interval I can belong to any of the following five sub-classes:

- $\mathcal{P}_1 := \{I : I \subseteq I_0, I \in D_M\}$.
- $\mathcal{P}_2 := \{I : I \supseteq I_0, I \in D_M\}$.

- $\mathcal{P}_3 := \{I : I \cap I_0 = \phi, I \in D_M\}$.
- $\mathcal{P}_4 := \{(a,b) : a < L < b < R, I = (a,b) \in D_M\}$.
- $\mathcal{P}_5 := \{(a,b) : L < a < R < b, I = (a,b) \in D_M\}$.

Subsequently, we detail the analysis for the relatively harder case $\mathcal{P}_4$. The arguments for the other cases are similar. Observe that:

$$\begin{aligned}
\mathbb{P}(\hat{I} \in \mathcal{P}_4) &\leq \mathbb{P}(\min_{I:I\in\mathcal{P}_4} V_I - V_{I_0} \leq 0) \\
&\leq \mathbb{P}\Big(\max_{I:I\in\mathcal{P}_4, x_1+x_3>M} \frac{S^\varepsilon_1 - S^\varepsilon_3}{x_1+x_3} \geq \rho\tilde{d}\Big) \\
&\leq \sum_{j=M+1}^{\infty} \inf_{\theta\geq 0} \mathbb{P}\Big(\max_{a,b:a<L<b<R:L-a+R-b=j} \exp(\theta(S^\varepsilon_{[a,L]} - S^\varepsilon_{[b,R]})) \geq \exp(\theta\rho\tilde{d}j)\Big) \\
&\leq \sum_{j=M+1}^{\infty} \inf_{\theta\geq 0} \exp(-\theta\rho\tilde{d}j)\mathbb{E}\Big[\max_{a,b:a<L<b<R:L-a+R-b=j} \exp(\theta(S^\varepsilon_{[a,L]} - S^\varepsilon_{[b,R]}))\Big] \\
&\leq \sum_{j=M+1}^{\infty} \inf_{\theta\geq 0} \exp(-\theta\rho\tilde{d}j)\mathbb{E}\Big[\max_{a,b:a\in\{L-j+1,\cdots,L\},b\in\{(R-j+1)\vee L,\cdots,R\}} \exp(\theta(S^\varepsilon_{[a,L]} - S^\varepsilon_{[b,R]}))\Big]
\end{aligned} \tag{11}$$

For $j \in [n]$, let $\mathcal{F}_j := \sigma(\{(\omega_{s-1}, \zeta_s) : s < j\})$. Write

$$\begin{aligned}
&\mathbb{E}\Big[\max_{a,b:a\in\{L-j+1,\cdots,L\},b\in\{(R-j+1)\vee L,\cdots,R\}} \exp(\theta(S^\varepsilon_{[a,L]} - S^\varepsilon_{[b,R]}))\Big] \\
&= \mathbb{E}\Big[\max_{a:a\in\{L-j+1,\cdots,L\}} \exp(\theta S^\varepsilon_{[a,L]})\mathbb{E}\Big[\max_{b\in\{(R-j+1)\vee L,\cdots,R\}} \exp(-\theta S^\varepsilon_{[b,R]}) \mid \mathcal{F}_{(R-j)\vee L}\Big]\Big] \\
&\leq \mathbb{E}\Big[\max_{a:a\in\{L-j+1,\cdots,L\}} \exp(\theta S^\varepsilon_{[a,L]})\sqrt{\mathbb{E}[\exp(-2\theta S^\varepsilon_{[(R-j+1)\vee L,R]}) \mid \mathcal{F}_{(R-j)\vee L}]} \\
&\qquad \sqrt{\mathbb{E}\Big[\max_{b\in\{(R-j+1)\vee L,\cdots,R\}} \exp(2\theta S^\varepsilon_{[(R-j+1)\vee L,b]}) \mid \mathcal{F}_{(R-j)\vee L}\Big]}\Big],
\end{aligned} \tag{12}$$

where, (12) follows from Cauchy-Schwartz inequality. Now, note that, by construction of ε_t, conditional on $\mathcal{F}_{(R-j)\vee L}$, ε_t is a martingale difference sequence adapted to $\sigma(\{(\omega_{s-1}, \zeta_s) : (R-j+1)\vee L \leq s \leq t\})$. Since $x \mapsto \exp(2\theta x)$ is convex, hence $\exp(2\theta S^\varepsilon_{[(R-j+1)\vee L,b]}), b \in \{(R-j+$

$1) \vee L, \ldots, R\}$ is a sub-martingale sequence. Consequently, Doob's maximal inequality (Hall & Heyde 1980) applies. Further sequential conditioning yields the following series of inequalities.

$$\begin{aligned}
&\mathbb{E}\Big[\max_{b\in\{(R-j+1)\vee L,\cdots,R\}} \exp(2\theta S^{\varepsilon}_{[(R-j+1)\vee L,b]}) \mid \mathcal{F}_{(R-j)\vee L}\Big] \\
&\leq 4\mathbb{E}[\exp(2\theta S^{\varepsilon}_{[(R-j+1)\vee L,R]}) \mid \mathcal{F}_{(R-j)\vee L}] \\
&\leq 4\mathbb{E}[\exp(2\theta S^{\varepsilon}_{[(R-j+1)\vee L,R-1]})\mathbb{E}[\exp(2\theta\varepsilon_R)|\mathcal{F}_{R-1}] \mid \mathcal{F}_{(R-j)\vee L}] \\
&\leq 4\sup_P \mathbb{E}_{1,P}[\exp(2\theta\varepsilon)]\mathbb{E}[\exp(2\theta S^{\varepsilon}_{[(R-j+1)\vee L,R-1]}) \mid \mathcal{F}_{(R-j)\vee L}] \\
&\leq 4\Big(\sup_P \mathbb{E}_{1,P}[\exp(2\theta\varepsilon)]\Big)^j. \qquad (13)
\end{aligned}$$

Proceeding along similar lines, we obtain

$$\mathbb{E}[\exp(-2\theta S^{\varepsilon}_{[(R-j+1)\vee L,R]}) \mid \mathcal{F}_{(R-j)\vee L}] \leq 4\Big(\sup_P \mathbb{E}_{1,P}[\exp(-2\theta\varepsilon)]\Big)^j, \qquad (14)$$

and

$$\mathbb{E}[\max_{a:a\in\{L-j+1,\cdots,L\}} \exp(\theta S^{\varepsilon}_{[a,L]})] \leq 4\Big(\mathbb{E}_0[\exp(\theta\varepsilon)]\Big)^j. \qquad (15)$$

Combining (13)-(15) and plugging them in (12) and (11), one obtains

$$\mathbb{P}(I \in \mathcal{P}_4) \leq 16 \sum_{j=M+1}^{\infty} \inf_{\theta\geq 0} \left(\exp(-\theta\rho\tilde{d})\mathbb{E}_0[\exp(\theta\varepsilon)]\sqrt{\sup_P \mathbb{E}_{1,P}[\exp(2\theta\varepsilon)] \sup_P \mathbb{E}_{1,P}[\exp(-2\theta\varepsilon)]} \right)^j. \qquad (16)$$

To deliver the coup de grâce of our argument, we are required to bound (16). To that end, define $\phi : \mathbb{R}_+ \to \mathbb{R}$ as

$$\phi(\theta) = -\theta\rho\tilde{d} + \log \mathbb{E}_0[\exp(\theta\varepsilon)] + 2^{-1} \log \sup_P \mathbb{E}_{1,P}[\exp(2\theta\varepsilon)] + 2^{-1} \log \sup_P \mathbb{E}_{1,P}[\exp(-2\theta\varepsilon)].$$

By definition of ϕ, $\mathbb{P}(I \in \mathcal{P}_4) \leq \sum_{j=M+1}^{\infty} \inf_{\theta \geq 0} \exp(j\phi(\theta))$. Moreover, for $\lambda \in (0,1)$, $\theta_1, \theta_2 \in \mathbb{R}_+$, Hölder's inequality produces

$$\log \sup_P \mathbb{E}_{1,P}[\exp(2(\lambda\theta_1 + (1-\lambda)\theta_2)\varepsilon)] \leq \sup_P \Big(\lambda \log \mathbb{E}_{1,P}[\exp(2\theta_1\varepsilon)] + (1-\lambda) \log \mathbb{E}_{1,P}[\exp(2\theta_2\varepsilon)]\Big). \tag{17}$$

Similar arguments for $\log \mathbb{E}_0[\exp(\theta\varepsilon)]$ and $\log \sup_P \mathbb{E}_{1,P}[\exp(-2\theta\varepsilon)]$ show that ϕ, being a linear combination of convex functions with non-negative weights (note that $-\theta\rho\tilde{d}$ is linear) , is itself convex.

Let $f : \mathbb{R} \otimes \mathbb{R}^{|W|} \mapsto \mathbb{R}$ be given by $f(\theta, P) = \log \mathbb{E}_0[\exp(2\theta\varepsilon)] + \log \mathbb{E}_{1,P}[\exp(2\theta\varepsilon)]$. Recalling that $f^{(1)}(\theta, w) = \frac{\partial}{\partial\theta} f(\theta, w)$, observe that

$$f^{(1)}(0, P) = 0 \text{ for any } P \in \mathcal{P}, \tag{18}$$

since $\mathbb{E}_0[\varepsilon] = \mathbb{E}_{1,P}[\varepsilon] = 0$. Therefore, noting that $\mathcal{P}$ is a compact subset of the $|W|$-dimensional simplex, in light of $\sup_{P\in\mathcal{P}} \mathbb{E}[\exp(-\eta|\varepsilon|)] \leq \sup_{P\in\mathcal{P}} \mathbb{E}[\exp(\eta|\varepsilon|)] < \infty$, Danskin's Theorem (Danskin 1967) entails

$$\frac{\partial}{\partial\theta} \sup_{P\in\mathcal{P}} f(\theta, P)\Big|_{\theta\downarrow 0} = \sup_{P\in\mathcal{P}} f^{(1)}(0, P) = 0,$$

where in the second inequality we use that $f(0, P) = 0$ for any $P \in \mathcal{P}$, and the third equality follows from (18). Similarly, $\frac{\partial}{\partial\theta} \sup_{P\in\mathcal{P}} f(\theta, P)\Big|_{\theta\uparrow 0} = -\inf_{P\in\mathcal{P}} f^{(1)}(0, P) = 0$. Therefore, $\phi'(0) = -\rho\tilde{d} < 0$. On the other hand, since $\min\{\mathrm{Var}_0(\varepsilon), \sup_P \mathrm{Var}_{1,P}(\varepsilon)\} > 0$, hence $\phi(\theta) \to \infty$ as $\theta \uparrow \infty$. In conjunction with ϕ being convex, there must exist $\kappa \in (0,1)$ such that $\log \kappa := \inf_{\theta\geq 0} \phi(\theta)$.

Consequently, from (16), one obtains,

$$\mathbb{P}(I \in \mathcal{P}_4) \leq 16 \sum_{j=M+1}^{\infty} \kappa^j = O(\kappa^M).$$

Suppose $\delta \in (0,1)$ be given. A choice of $M > \frac{\log 1/\varepsilon}{\log 1/\kappa}$ ensures that $\mathbb{P}(I \in \mathcal{P}_4) < \delta$. This completes the proof. □

Finally, Theorem 3.1 is proved by invoking Theorem D.1 and Proposition 2.

We can further sharpen the $O((\rho\tilde{d})^{-1})$ rate in Theorem 3.1 to $O((\rho\tilde{d})^{-2})$ by assuming a mild condition: local sub-Gaussianity of the pivot statistics. The following result also trivially follows from Theorem D.1 and Proposition 2, but is separately stated to highlight its importance.

Lemma D.2. *Grant the assumptions of Theorem 3.1. If*

$$\max\{\mathbb{E}_0[\exp(r|\varepsilon|)], \sup_{P \in \mathcal{P}} \mathbb{E}_{1,P}[\exp(r|\varepsilon|)]\} \leq \exp(r^2/2),$$

for all $r \in [0, \eta]$, then choosing $\rho > 0$ such that $\rho\tilde{d} < \frac{5}{2}\eta$, then $|\hat{I}\Delta I_0| = O_{\mathbb{P}}\Big((\rho\tilde{d})^{-2}\Big)$.

D.2 Proof of Theorem 4.1

For convenience, we first re-state the theorem.

Theorem D.3. *Assume that the null distribution of the pivot statistics is absolutely continuous with respect to the Lebesgue measure. Let the number of watermarked intervals K be bounded, and Assumption 4.1 be granted for the watermarked intervals $I_k, k \in [K]$. Fix $\alpha \in (0,1)$, and recall the quantities defined in* `WISER` *described in Figure 2. Suppose that $\mathbb{E}_0[|X - \mu_0|^p] < \infty$ for some $p \geq 2$, and let the block length $b = b_n$ satisfy $b_n = O(n^{\upsilon})$, and $b_n/n^{1/p} \to \infty$, where $\upsilon > 1/p$ is same as in Assumption 4.1. Moreover, suppose the threshold $\mathcal{Q} = \mathcal{Q}_n$ is selected so that*

$\mathbb{P}_0(\max_{1\le k\le\lceil n/b\rceil} S_k > \mathcal{Q}) = \alpha$. *Finally, assume* $d \ge c$ *for some constant* $c > 0$*, and*

$$\sup_{P\in\mathcal{P}} \mathbb{E}_{1,P}[X] < \infty, \tag{19}$$

and assume there exists $\tau > 0$ *such that*

$$\kappa := \inf_{\theta\ge 0} \theta(\mu_0 + \tau d) + \log \sup_{P} \mathbb{E}_{1,p}[\exp(-\theta X)] < 0. \tag{20}$$

Then, given $\varepsilon > 0$ *and* $d \ge c$ *for some constant* $c > 0$*, under the assumptions of Theorem 3.1, there exist* $M_\varepsilon \in \mathbb{R}_+$*, independent of* n, K*, and* d*, and* $\rho > 0$*, such that* `WISER` *applied with hyper-parameters* b *and* ρ *satisfies*

$$\liminf_{n\to\infty} \mathbb{P}\Big(\hat{K} = K,\ \max_{k\in[K]} |\hat{I}_k \Delta I_k| < M_\varepsilon d^{-1}\Big) \ge 1 - \varepsilon. \tag{21}$$

Let $\widetilde{\mathcal{B}} = \{1 \le k \le \lceil n/b\rceil : B_k \subseteq I_j \text{ for some } j \in [K]\}$. Our proof proceeds through a series of arguments, each carefully orchestrated to establish the validity of the corresponding steps of our algorithm. We comment that subsequently, all statements involving n but without a limit attached to it are meant to be considered for all sufficiently large values of n.

Step 1: Validity of first stage thresholding.

In this step, we show that

$$\mathbb{P}(\min_{k\in\widetilde{\mathcal{B}}} S_k > \mathcal{Q}) \to 1, \text{ as } n \to \infty. \tag{22}$$

To begin with, note that

$$\limsup_{n\to\infty} \max_{k\in\widetilde{\mathcal{B}}} \mathbb{P}(S_k \le \mathcal{Q}_n)^{1/b} \le \limsup_{n\to\infty} \inf_{\theta\ge 0} \exp(\theta \mathcal{Q}_n b_n^{-1} + \log \sup_{P\in\mathcal{P}} \mathbb{E}_{1,P}[\exp(-\theta X)])$$

$$\le \inf_{\theta\ge 0} \exp(\theta\mu_0 + \log \sup_{P\in\mathcal{P}} \mathbb{E}_{1,P}[\exp(-\theta X)]) \tag{23}$$

$$\le \exp(\kappa) < 1, \tag{24}$$

where (23) is obtained through an application of Proposition 3, and (24) follows from (20). Since $\kappa < 0$, one has $\frac{n}{b}\exp(\kappa b) \to 0$ as $n \to \infty$, and consequently

$$\mathbb{P}(\min_{k\in\widetilde{\mathcal{B}}} S_k \le \mathcal{Q}) \le \frac{n}{b} \max_{k\in\widetilde{\mathcal{B}}} \mathbb{P}(S_k \le \mathcal{Q}_n) \to 0, \text{ as } n \to \infty,$$

thereby establishing (22).

Step 2. Estimation of the number of watermarked regions through the set M.

Recall M from the Step 1 of Subroutine `Refined_Local_Search` in Algorithm 3. In this step of our proof, we will prove $\mathbb{P}(\hat{K} = K) \to 1$, which will also imply that $|M|$ is even with probability approaching 1. Therefore, we may be excused for assuming that $|M|$ is even.

Let $C_1, \ldots, C_{\hat{K}}$ be the disjoint set of intervals in M, with $C_j = [(s_{2j-1} - 1)b + 1, s_{2j}b]$. Note that for each $k \in \mathcal{B}$ such that $S_k > \mathcal{Q}$, $B_k \subseteq C_j$ for some j. Let $\widetilde{\mathcal{B}}_j = \{k \in \widetilde{\mathcal{B}} : B_k \subseteq I_j\}$, $j \in [K]$. We remark that

Clearly, $\widetilde{\mathcal{B}} = \cup_{j=1}^{K} \widetilde{\mathcal{B}}_j$, and $\widetilde{\mathcal{B}}_j$ are disjoint. Therefore, in light of the construction of M from blocks surpassing the threshold $\mathcal{Q}$, it follows,

$$\mathbb{P}(\min_{k\in\widetilde{\mathcal{B}}} S_k > \mathcal{Q}) = \mathbb{P}(\min_{j\in[K]} \min_{k\in\widetilde{\mathcal{B}}_j} S_k > \mathcal{Q}) \le \mathbb{P}(\text{for each } j \in [K], \text{ there exists } i_j \in [\hat{K}] \text{ such that } \widetilde{\mathcal{B}}_j \subseteq C_{i_j}),$$

which implies, in light of (22),

$$\mathbb{P}(A_n) \to 1, \text{ as } n \to \infty, \text{ where, } A_n := \{\text{for each } j \in [K], \text{ there exists } i_j \in [\hat{K}] \text{ such that } \widetilde{\mathcal{B}}_j \subseteq C_{i_j}\}. \tag{25}$$

It is crucial to note that since both $\widetilde{\mathcal{B}}_j$ and C_j's are defined to occur from left-to-right and since C_j's are connected intervals, under the event A_n it also holds that $i_1 \leq i_2 \leq \ldots \leq i_K$. At this stage, the relationship between $\hat{K}$ and K is still not entirely clear. Subsequently, we will show that under the event A_n, the mapping $j \mapsto i_j$ is injective, establishing that $\hat{K} \geq K$ with high probability. To that end, suppose there exists $k_1 < k_2 \in [K]$ such that $i_{k_1} = i_{k_2}$. Since $C_{i_{k_1}}$ is a connected interval, $i_{k_1} = i_{k_2}$ implies that that $i_{k_1} = i_{k_1+1}$. Let $\mathbb{P}_{E,F}(\cdot) = \mathbb{P}(\cdot \cap E \cap F)$ for any events E, F. Consider the following series of inequalities.

$$\begin{aligned}
&\mathbb{P}_{A_n}(\text{There exists } k \in [K-1] \text{ such that } C_{i_k} = C_{i_{k+1}}) \\
&\leq \mathbb{P}_{A_n}(\text{There exists } k \in [K-1] \text{ such that } (I_{k,R}, I_{k+1,L}) \subseteq C_{i_k}) \\
&\leq \mathbb{P}_{A_n}(\text{There exists } k \text{ such that } \min_{l \in (\lceil I_{k,R}/b \rceil, \lfloor I_{k+1,L}/b \rfloor)} S_l > \mathcal{Q}) \\
&\leq \mathbb{P}_0(\sum_{k=1}^{n/b} I\{S_k > \mathcal{Q}\} \geq C_0\sqrt{\log n}),
\end{aligned} \tag{26}$$

where the $\mathbb{P}_0$ in final inequality appears since for $l \in (\lceil I_{k,R}/b \rceil, \lfloor I_{k+1,L}/b \rfloor)$, the region B_l is unwatermarked; the $\sqrt{\log n}$ appears by invoking Assumption 4.1 and noting that $b^{-1}(I_{k+1,L} - I_{k,R}) \geq C_0\sqrt{\log n}$. An application of Proposition 4 to (26) entails, in view of (25), that,

$$\mathbb{P}_{A_n}(\bar{B}_n) \to 1, \text{ as } n \to \infty, \text{ where } \bar{B}_n = \{C_{i_k} \text{ and } C_{i_s} \text{ are disjoint if } i_k \neq i_s\}.$$

Clearly, this implies that $\mathbb{P}_{A_n}(\hat{K} \geq K) \to 1$ as $n \to \infty$, which also produces $\mathbb{P}(\hat{K} \geq K) \to 1$ as $n \to \infty$. On the other hand, if $\hat{K} > K$, then under the event $A_n \cap \bar{B}_n$, there exists $j \in [\hat{K}]$ such

that C_j and $\cup_{s\in\mathcal{B}}B_s$ are disjoint. Consequently, it must be true that $|C_j \cap (\cup_{k=1}^K I_j)| \leq b$. Note that, by construction of C_j's in `WISER`, $|C_j| \geq cb\sqrt{\log n}$. Therefore it must be true that there are at least $2^{-1}c\sqrt{\log n}$ many s's such that $B_s \cap C_j \cap (\cup_{k=1}^K I_j) = \phi$, and $S_s > \mathcal{Q}$. Hence it follows from Proposition 4 that

$$\mathbb{P}_{A_n,\bar{B}_n}(\hat{K} > K) \to 0, \text{ as } n \to \infty,$$

which immediately implies that

$$\mathbb{P}(\hat{K} = K) \to 1 \text{ as } n \to \infty. \tag{27}$$

Step 3. Choice of $\tilde{d}$ and ρ.

Recall $\tilde{d}$ from Step 5 of Subroutine `Refined_Local_Search` in Algorithm 3. In this step, we establish that there exists $\rho > 0$, such that $d > 2\rho\tilde{d}$ with high probability. In conjunction to $\tilde{d}$, also define

$$d^\dagger = \frac{\sum_{j=1}^K \sum_{s\in I_j}(X_s - \mu_0)}{\sum_{j=1}^K |I_j|}.$$

Let the event $\{\hat{K} = K\}$ be denoted as E_n. Under E_n, by construction of D_j, $\mathbb{P}_{A_n,B_n,E_n}(I_j \subseteq D_j \text{ for all } j \in [K]) \to 1$ as $n \to \infty$. Call the latter event as F_n. Observe that under $E_n \cap F_n$, it holds

$$\sum_{j=1}^K |I_j| + 2Cb\log n \geq \sum_{j=1}^{\hat{K}} |D_j| \geq \sum_{j=1}^K |I_j| + Cb\log n \tag{28}$$

for some $C > 0$. Therefore, under the same event, it follows

$$\tilde{d} \leq d^\dagger \frac{\sum_{j=1}^K |I_j|}{\sum_{j=1}^K |I_j| + Cb\log n} + \frac{\sum_{s\in\cup_j(I_j^c\cap D_j)}(X_s - \mu_0)}{\sum_{j=1}^K |I_j| + Cb\log n}. \tag{29}$$

We first tackle the second term in the upper-bound in (29). Let $D_j^\dagger = [(I_{j,L} - \lfloor Cb\log^{3/2} n\rfloor) \vee$

$1, (I_{j,R} + \lfloor Cb\log^{3/2} n \rfloor) \wedge n]$. Again, by construction of D_j as well as from Assumption 4.1, for all sufficiently large n it follows

$$\mathbb{P}_{A_n,\bar{B}_n,E_n}(D_j \subseteq D_j^\dagger, D_i^\dagger \cap D_j^\dagger = \phi \text{ for } i \neq j) \to 1.$$

Call the above event as G_n. Fix $\varepsilon > 0$, and consider the following implications.

$$\begin{aligned}
&\mathbb{P}_{A_n,\bar{B}_n,E_n}\left(\frac{\sum_{s\in\cup_j(I_j^c\cap D_j)}(X_s-\mu_0)}{\sum_{j=1}^K|I_j|+Cb\log n} > \varepsilon\right) \\
&\leq \mathbb{P}_{A_n,\bar{B}_n,E_n,G_n}\left(\frac{\sum_{s\in\cup_j(I_j^c\cap D_j^\dagger)}|X_s-\mu_0|}{\sum_{j=1}^K|I_j|+Cb\log n} > \varepsilon\right) + o(1) \\
&\leq \mathbb{P}\left(\frac{\sum_{s\in\cup_j(I_j^c\cap D_j^\dagger)}|X_s-\mu_0|}{\sum_{j=1}^K|I_j|+Cb\log n} > \varepsilon\right) + o(1) \\
&\leq \frac{O(b\log^{3/2} n)}{\varepsilon^2 n\log^2 n} + o(1) = o(1),
\end{aligned} \tag{30}$$

where the inequality in the final assertion follows from $|\cup_{j=1}^K (I_j^c \cap D_j^\dagger)| \lesssim b\log^{3/2} n$. Therefore, (29) and (30) jointly yields

$$\mathbb{P}_{A_n,\bar{B}_n,E_n,F_n}(\tilde{d} \leq 2d^\dagger) \to 1, \text{ as } n \to \infty. \tag{31}$$

Next, we focus on controlling $d^\dagger$ by d. To that end, we resort to an argument through moment generating functions. On one hand, (20) entails

$$\mathbb{P}(d^\dagger \leq \tau d) \leq \inf_{\theta\geq 0}\Big(\exp(\theta(\mu_0+\tau d) + \log\sup_{P\in\mathcal{P}}\mathbb{E}_{1,P}[\exp(-\theta X)])\Big)^{\sum_{j=1}^K|I_j|} \leq \exp(\kappa\sum_{j=1}^K|I_j|) \to 0. \tag{32}$$

On the other hand, in light of (19) and $d \geq c$, choose

$$\nu > \frac{\sup_{P\in\mathcal{P}} \mathbb{E}_{1,P}[X] - \mu_0}{c} \vee \frac{\tau}{4},$$

and write:

$$\mathbb{P}(d^\dagger \geq 2\nu d) \leq \inf_{\theta\geq 0} \left(\exp(-2\theta\nu c + \log \sup_{P\in\mathcal{P}} \mathbb{E}_{1,P}[\exp(\theta(X-\mu_0))] \right)^{\sum_{j=1}^K |I_j|}. \tag{33}$$

Echoing the argument in the proof of Theorem 3.1, define

$$g(\theta, P\,; c) = -2\theta\nu c + \log \mathbb{E}_{1,P}[\exp(\theta(X-\mu_0))],\ \widetilde{g}(\theta\,; c) = \sup_{P\in\mathcal{P}} g(\theta, P).$$

Since $\mathcal{P}$ is compact and $\sup_{P\in\mathcal{P}} \mathbb{E}_{1,P}[\exp(\eta|X-\mu_0|] < \infty$, Danskin's Theorem (Danskin 1967) applies and produces

$$\widetilde{g}_+^{(1)}(0, P\,; c) = \frac{\partial}{\partial\theta} \sup_{P\in\mathcal{P}} g(\theta, P\,; c)\Big|_{\theta\downarrow 0} = \sup_{P\in\mathcal{P}} g^{(1)}(0, P\,; c) = -2\nu d + \sup_{P\in\mathcal{P}} \mathbb{E}_{1,P}[X-\mu_0] \leq -\nu d < 0, \tag{34}$$

where the final inequality is derived via (19). Moreover, similar to (17) it can be argued that $\widetilde{g}(\theta)$ is convex in θ. Finally, since $\widetilde{g}(0\,; c) = 0$, (34) coupled with its convexity implies that $\varphi(c) := \inf_{\theta\geq 0} \widetilde{g}(\theta\,; c) < 0$. In view of this, (33) results in

$$\mathbb{P}(d^\dagger \geq 2\nu d) \leq \exp(-\varphi(c) \sum_{j=1}^K |I_j|) \to 0 \text{ as } n \to \infty, \tag{35}$$

where the limiting assertion is due to $\sum_{j=1}^K |I_j| \geq c\sqrt{n}$. Finally, (31) and (35) jointly indicates that

$$\mathbb{P}_{A_n,B_n,E_n,F_n}(\tilde{d} \leq 4\nu d) \to 1, \text{ as } n \to \infty. \tag{36}$$

Subsequently, we choose $\rho = (8\nu)^{-1}$. In conclusion to this step, (30) along with (32) establishes

$$\mathbb{P}_{A_n,B_n,E_n,F_n}(G_n) \to 1 \text{ as } n \to \infty,\ G_n := \{\tau d \leq \tilde{d} \leq 4\nu d\}.$$

Step 4. Localization of watermarked intervals.

In this step, we establish the validity of our localized estimates $\hat{I}_j$. In Step 3, we argued that

$$\mathbb{P}_{A_n,B_n,E_n}(I_j \subseteq D_j \subseteq D_j^{\dagger} \text{ for each } j \in [K]) \to 1 \text{ as } n \to \infty.$$

Call the above event as $\tilde{F}_n$. Under $\tilde{F}_n$, it is immediate that

$$\hat{I}_j(\tilde{d}) = \arg \min_{s \in L_j, t \in R_j} \sum_{k \in D_j \setminus [s,t]} (X_k - \mu_0 - \rho\tilde{d}) = \arg \min_{s \in L_j, t \in R_j} \sum_{k \in D_j^{\dagger} \setminus [s,t]} (X_k - \mu_0 - \rho\tilde{d}),$$

since the operator $\sum_{k \in D_j^{\dagger} \setminus [s,t]}$ can be decomposed into $\sum_{k \in D_j \setminus [s,t]} + \sum_{k \in D_j^{\dagger} \setminus D_j}$.

We proceed towards applying Theorem 3.1 to $\hat{I}_j(\tilde{d})$. However, note that $\tilde{d}$ is a random quantity, so special care must be accorded to its treatment. To that end, define

$$\hat{I}_j(\sigma) = \arg \min_{s \in L_j, t \in R_j} \sum_{k \in D_j^{\dagger} \setminus [s,t]} (X_k - \mu_0 - \rho\sigma),\ \sigma \in [\tau d, 4\nu d].$$

Fix $j \in [K]$. For $M > 0$, let $D_M := \{I : |I \Delta I_j| > M\}$. For a candidate interval $I = [s,t]$, let

$$\tilde{V}_I(\sigma) = \sum_{k \in D_j^{\dagger} \setminus [s,t]} (X_k - \mu_0 - \rho\sigma).$$

Clearly, by definition of G_n,

$$\mathbb{P}_{A_n,B_n,E_n,\tilde{F}_n,G_n}(|\hat{I}_j(\tilde{d}) \,\Delta\, I_n| > M)$$

$$\leq \mathbb{P}_{A_n,B_n,E_n,\tilde{F}_n,G_n}\Big(\sup_{\sigma\in[\tau d,4\nu d]} |\hat{I}_j(\sigma)\ \Delta\ I_n| > M\Big)$$

$$\leq \mathbb{P}_{A_n,B_n,E_n,\tilde{F}_n,G_n}\Big(\text{There exists } \sigma\in[\tau d,4\nu d] \text{ such that } \inf_{s\in L_j,t\in R_j,I\in D_M} \widetilde{V}_I(\sigma) < \widetilde{V}_{I_j}(\sigma)\Big)$$

$$\leq \mathbb{P}_{A_n,B_n,E_n,\tilde{F}_n,G_n}\Big(\text{There exists } \sigma\in[\tau d,4\nu d] \text{ such that } \inf_{I\in D_M} \widetilde{V}_I(\sigma) < \widetilde{V}_{I_j}(\sigma)\Big)$$

$$\leq \mathbb{P}_{A_n,B_n,E_n,\tilde{F}_n,G_n}\Big(\max_{I:x_1+x_3>M} \frac{S_1^\varepsilon - S_3^\varepsilon}{x_1+x_3} > (\frac{1}{2}\wedge\frac{\tau}{8\nu})d\Big) \tag{37}$$

$$\leq \mathbb{P}_{A_n,B_n,E_n,\tilde{F}_n,G_n}\Big(\max_{I:x_1+x_3>M} \frac{S_1^\varepsilon - S_3^\varepsilon}{x_1+x_3} > \frac{\tau}{8\nu}d\Big) \tag{38}$$

$$\leq \mathbb{P}\Big(\max_{I:x_1+x_3>M} \frac{S_1^\varepsilon - S_3^\varepsilon}{x_1+x_3} > \frac{\tau}{8\nu}d\Big). \tag{39}$$

Here, (37) follows by recalling the notations in the proof of Theorem 3.1 and following the arguments (9)-(10) after observing $\sigma\in[\tau d,4\nu d]$ implies $d-(8\nu)^{-1}\sigma\geq\frac{d}{2}$. Moreover, (38) also follows from $4\nu d\geq\sigma\geq\tau d$. Finally, (39) is derived from $\mathbb{P}(A\cap B)\leq\mathbb{P}(A)$; in particular, arguments of Theorem 3.1 can be followed verbatim to obtain that

$$\mathbb{P}\Big(\max_{I:x_1+x_3>M} \frac{S_1^\varepsilon - S_3^\varepsilon}{x_1+x_3} > \frac{\tau}{8\nu}d\Big) \leq \xi^M \text{ for some } \xi<1.$$

Note that in the above assertion we have used the fact that $d\geq c$ to decouple ξ from d. Given arbitrary $\varepsilon>0$, M_ε can be chosen to ensure $\xi^{M_\varepsilon}<\varepsilon$, and through κ, this choice of M_ε solely depends on the constants ν, τ, c, and μ_0, apart from the quantity ε. Therefore, in view of the number of watermarked intervals $K=O(1)$, we obtain that there exists M_ε independent of n, K and d such that

$$\begin{aligned} &\mathbb{P}_{A_n,B_n,E_n,\tilde{F}_n,G_n}(|\hat{I}_j(\tilde{d})\ \Delta\ I_n| > M_\varepsilon \text{ for } j\in[K]) \leq \varepsilon \\ \implies &\liminf_{n\to\infty} \mathbb{P}_{A_n,B_n,E_n,\tilde{F}_n,G_n}(|\hat{I}_j(\tilde{d})\ \Delta\ I_n| \leq M_\varepsilon \text{ for } j\in[K]) \geq 1-\varepsilon, \end{aligned} \tag{40}$$

where in (40) we invoke

$$\lim_{n\to\infty} \mathbb{P}(A_n \cap B_n \cap E_n \cap \tilde{F}_n \cap G_n) = 1.$$

Recalling that $E_n = \{\hat{K} = K\}$ completes the proof.

D.3 Additional propositions

Firstly we provide a formal proof that the pivot statistics corresponding to un-watermarked tokens are i.i.d., a fundamental fact behind the construction and validity of our algorithm.

Proof of Lemma 2.2. Let $t \in S$. Since ω_t and ζ_t are independent conditional on $\omega_{1:(t-1)}$, hence $\mathcal{L}(Y_t)|\omega_{1:t} \overset{d}{=} \mathcal{L}(Y)$. Hence, $\{Y_t\}_{t\in S}$ are identically distributed since given $\omega_{1:t}$, the distribution of Y_t is solely a function of the key ζ_t that are i.i.d.. Moreover, for $s < t \in S$, even if there is a watermarked region $I_k \subset (s,t)$, ζ_s and ω_l are independent for all $l \in (s,t]$. In view of the fact that conditional on $\omega_{1:s}$, Y_s and Y_t are completely determined by ζ_s and $(w_{s+1:t}, \zeta_{s+1:t})$ respectively, we deduce that Y_s and Y_t are independent conditional on $\omega_{1:s}$. Hence, for two Borel sets A and B,

$$\begin{aligned}\mathbb{P}(Y_s \in A, Y_t \in B) =& \mathbb{E}[\mathbb{P}(Y_s \in A|\omega_{1:s})\mathbb{E}[\mathbb{P}(Y_t \in B|\omega_{1:t})|\omega_{1:s}]] \\ =& \mathbb{P}(Y_s \in A)\mathbb{P}(Y_t \in B) \text{ (from Definition 2.1).}\end{aligned}$$

This completes the proof. □

Next, we collect the additional results that we have used in our theoretical arguments. The proofs are provided subsequently.

Proposition 1. *Let $h(x) = -\log(1-x)$, and suppose $\mathcal{P}_\Delta := \{\max_{w\in\mathcal{W}} P_w \leq 1-\Delta\}$ for some*

fixed $\Delta > 0$. Then it follows that

$$\inf_{P \in \mathcal{P}_\Delta} \mathbb{E}_{1,P}[h(Y)] \geq \sum_{n=1}^{\infty} \left(\frac{1}{n} - \lfloor \frac{1}{1-\Delta} \rfloor \frac{(1-\Delta)^2}{1+n(1-\Delta)} - \frac{1-(1-\Delta)\lfloor \frac{1}{1-\Delta} \rfloor}{1+n(1-(1-\Delta)\lfloor \frac{1}{1-\Delta} \rfloor)} \right). \tag{41}$$

Proposition 2. *Consider $\tilde{d}$ and $\Psi(\cdot)$ from Theorem D.1. If there exists a constant $c > 0$ such that $d \geq c$, then*

$$(\sup_{\theta \geq 0} \{\theta \rho \tilde{d} - \Psi(\theta)\})^{-1} = O((\rho \tilde{d})^{-1}). \tag{42}$$

Recall ε from Theorem 3.1. Suppose we additionally have that

$$\max\{\mathbb{E}_0[\exp(r|\varepsilon|)], \sup_{P \in \mathcal{P}} \mathbb{E}_{1,P}[\exp(r|\varepsilon|)]\} \leq \exp(r^2/2) \text{ for all } r \in [0, \eta],$$

η being the same as in Theorem 3.1. Then, choosing $\rho > 0$ such that $\rho \tilde{d} < \frac{5}{2}\eta$, it holds that

$$(\sup_{\theta \geq 0} \{\theta \rho \tilde{d} - \Psi(\theta)\})^{-1} = O((\rho \tilde{d})^{-2}). \tag{43}$$

Proposition 3. *Let $\mathbb{E}_0[|X - \mu_0|^p] < \infty$ for some $\delta > 0$. Let $\mathcal{Q}, b$ be selected as in Theorem 4.1. Then it follows that $Q/b \to \mu_0$ as $n \to \infty$.*

Proposition 4. *Let X_i be i.i.d. with mean μ_0, and let B_k and S_k be defined as in Steps 2 and 3 of* `WISER` *in Figure 2. Then it follows that*

$$\mathbb{P}_0 \left(\sum_{k=1}^{\lceil n/b \rceil} I\{S_k > \mathcal{Q}\} \geq C_0 \sqrt{\log n} \right) \to 0, \text{ as } n \to \infty,$$

where $\mathcal{Q}$ is defined as in Theorem 4.1.

Proof of Proposition 1. From Lemma 3.1 of Li, Ruan, Wang, Long & Su (2025*b*), it follows

$$\begin{aligned}
\mathbb{E}_{1,P}[h(X)] &= \sum_{w=1}^{|\mathcal{W}|} \int_0^1 x^{1/P_w - 1}\Big(-\log(1-x)\Big)\,\mathrm{d}x \\
&= \sum_{w=1}^{|\mathcal{W}|} \sum_{n=1}^{\infty} \int_0^1 \frac{x^{1/P_w - 1 + n}}{n}\,\mathrm{d}x \\
&= \sum_{w=1}^{|\mathcal{W}|} \sum_{n=0}^{\infty} \frac{1}{n(n+1/P_w)} \\
&= \sum_{n=1}^{\infty} \Big(\frac{1}{n} - \sum_{w=1}^{|\mathcal{W}|} \frac{P_w}{n+1/P_w}\Big) \\
&\geq \sum_{n=1}^{\infty} \left(\frac{1}{n} - \lfloor \frac{1}{1-\Delta} \rfloor \frac{(1-\Delta)^2}{1+n(1-\Delta)} - \frac{1-(1-\Delta)\lfloor \frac{1}{1-\Delta} \rfloor}{1+n(1-(1-\Delta)\lfloor \frac{1}{1-\Delta} \rfloor)}\right), \qquad (44)
\end{aligned}$$

where the final inequality follows from noting the convexity of $g : x \mapsto \sum_{i=1}^{d} \frac{x_i}{n+1/x_i}$, $\sum_{i=1}^{d} x_i = 1$, and noting that the optimum value of g on the set $\mathcal{P}_\Delta$ occurs at the extrema defined by

$$P_\Delta^\star = \Big(\underbrace{1-\Delta, \ldots, 1-\Delta}_{\lfloor \frac{1}{1-\Delta} \rfloor \text{ times}}, 1-(1-\Delta)\cdot \lfloor \tfrac{1}{1-\Delta} \rfloor, 0, \ldots\Big).$$

□

Proof of Proposition 2. Denote $\Lambda(x) := \sup_{\theta \geq 0}\{\theta\rho x - \Psi(\theta)\}$. Note that, an argument same as (18) shows that $\Psi_+'(0) = 0$, where $\Psi_+'(\cdot)$ denote the right derivative. Therefore, in light of $d \geq c$ for some constant $c > 0$, there exist $\theta_0 > 0$ such that $\frac{|\Psi(\theta)|}{\theta} \leq \frac{\rho\tilde{d}}{2}$ for all $\theta \in (0, \theta_0)$. Therefore,

$$\Lambda(\tilde{d}) \geq 2^{-1}\theta_0\rho\tilde{d} - 4^{-1}\theta_0\rho c \geq 4^{-1}\theta_0\rho\tilde{d},$$

which immediately implies (42). Moving on, we work with the additional assumption that

$\sup_{P\in\mathcal{P}} \mathbb{E}_{1,P}[\exp(r|\varepsilon|)] \leq \exp(r^2/2)$. This immediately implies that for all $\theta \in [0, \frac{\eta}{2}]$,

$$\max\{\log \sup_P \mathbb{E}_{1,P}[\exp(2\theta\varepsilon)], \log \sup_P \mathbb{E}_{1,P}[\exp(-2\theta\varepsilon)]\} \leq 2\theta^2.$$

Therefore, for all $\theta \in [0, \frac{\eta}{2}]$ it must hold that

$$\Psi(\theta) \leq \frac{5}{2}\theta^2.$$

Consequently, in light of $\rho\tilde{d} < \frac{5}{2}\eta$, one obtains,

$$\Lambda(x) \geq \sup_{\theta\in[0,\frac{\eta}{2}]} \{\theta\rho x - \frac{5}{2}\theta^2\} = \frac{\rho^2 x^2}{10},$$

which establishes (43). □

Proof of Proposition 3. Our proof has two key steps: firstly, we will prove that if there is no watermarking in the entire sequence, then

$$\max_{1\leq k\leq \lceil n/b\rceil} \frac{S_k}{b} \xrightarrow{\mathbb{P}} \mu_0. \tag{45}$$

Subsequently, we follow an argument similar to the proof of equation (29) in Li, Ruan, Wang, Long & Su (2025*b*), with crucial tweaks to accommodate the maximum over the block means. Let us first work towards (45). We note that a similar result (for the p-th moments) appears in Proposition E.2 in Deb et al. (2020) but without proof. For the sake of completion, we provide an independent proof of (45) without invoking the aforementioned result. Fix $\varepsilon > 0$. Note that

$$\mathbb{P}_0(\max_{1\leq k\leq \lceil n/b\rceil} b^{-1}(S_k - \mu_0) > \varepsilon) \leq \frac{n}{b}\mathbb{P}(b^{-1}(S_1 - \mu_0) > \varepsilon), \tag{46}$$

where for the last inequality we use that S_k's are i.i.d. under H_0, i.e. no watermarking. Moving on, we apply the Fuk-Nagaev inequality (Corollary 4, Fuk & Nagaev (1971)),

$$\mathbb{P}_0(b^{-1}(S_1-\mu_0)>\varepsilon) \le c_1\frac{b}{(b\varepsilon)^p}\mathbb{E}_0[|X-\mu|^p] + \exp(-c_2\frac{b\varepsilon^2}{\sigma^2}),\ \sigma^2 := \mathbb{E}_0[X^2], \tag{47}$$

where $c_1, c_2 > 0$ are constants depending solely on p. Note that $b^p \lesssim n$, and hence $\frac{n}{(b\varepsilon)^p} \to 0$ as $n \to \infty$. On the other hand, $nb^{-1}\exp(-c_{2,\delta}\frac{b\varepsilon^2}{\sigma^2}) \to 0$ as $n \to \infty$. Therefore, from (46) and (47), one obtains (45).

Now suppose that $\limsup_{n\to\infty} \mathcal{Q}/b > \mu_0$. Then there exists $\gamma > 0$ and a strictly increasing sequence $\{n_k\} \subseteq \mathbb{N}$ such that $\mathcal{Q}_{n_k}/b_{n_k} > \mu_0 + \gamma$ for all sufficiently large $k \in \mathbb{N}$. Since (45) implies that

$$\max_{1\le l\le \lceil n_k/b_{n_k}\rceil} \frac{S_{n_k}}{b_{n_k}} \xrightarrow{\mathbb{P}} \mu_0, \text{ as } k\to\infty,$$

therefore, there exists a strictly increasing sub-sequence $\{n_{k_r}\} \subseteq \{n_k\}$ such that

$$\max_{1\le l\le \lceil n_{k_r}/b_{n_{k_r}}\rceil} \frac{S_{n_{k_r}}}{b_{n_{k_r}}} \xrightarrow{\text{a.s.}} \mu_0, \text{ as } r\to\infty, \text{ and } \mathcal{Q}_{n_{k_r}}/b_{n_{k_r}} > \mu_0+\gamma \text{ for all sufficiently large } r.$$

Therefore, by the dominated convergence theorem,

$$\begin{aligned}\alpha = \lim_{r\to\infty}\mathbb{P}\left(\max_{1\le l\le \lceil n_{k_r}/b_{n_{k_r}}\rceil}\frac{S_{n_{k_r}}}{b_{n_{k_r}}} > \frac{\mathcal{Q}_{n_{k_r}}}{b_{n_{k_r}}}\right) &\le \lim_{r\to\infty}\mathbb{P}\left(\max_{1\le l\le \lceil n_{k_r}/b_{n_{k_r}}\rceil}\frac{S_{n_{k_r}}}{b_{n_{k_r}}} > \mu_0+\gamma\right)\\ &= \mathbb{P}(\mu_0 > \mu_0+\gamma) = 0,\end{aligned} \tag{48}$$

which is a contradiction. Hence, $\limsup_{n\to\infty}\mathcal{Q}/b \le \mu_0$. Very similarly one can show $\liminf_{n\to\infty}\mathcal{Q}/b \ge \mu_0$, which completes the proof. □

Proof of Proposition 4. Let $p_n = \mathbb{P}_0(S_k > \mathcal{Q})$. Clearly, by definition of $\mathcal{Q}$ it follows that

$$\alpha = \mathbb{P}_0(\max_k S_k > \mathcal{Q}) = 1 - (1 - p_n)^{\lceil n/b \rceil} \geq 1 - \exp(-c\sqrt{n}p_n),$$

which implies that $\sqrt{n}p_n = O(1)$. Note that $\sum_{k=1}^{\lceil n/b \rceil} I\{S_k > \mathcal{Q}\} \sim \text{Bin}(\lceil n/b \rceil, p_n)$. Therefore, using Chernoff bound, one obtains

$$\mathbb{P}_0(\sum_{k=1}^{\lceil n/b \rceil} I\{S_k > \mathcal{Q}\} \geq C_0\sqrt{\log n}) \leq \exp(-2^{-1}(1 + o(1))(\log\log n)\log n) \to 0,$$

which completes the proof. □