Journal of the Royal Statistical Society Series B: Statistical Methodology, , 1–34
doi:
Original Article

Optimal Watermark Localization in Mixed-Source Large Language Model Texts

Jose H. Blanchet,[1] T. Tony Cai,[2] Xiang Li,[2,*] Hao Liu,[1] Qi Long[2] and Weijie J. Su[2]

[1]Stanford University, USA and [2]University of Pennsylvania, USA

Author names are listed in alphabetical order. *Corresponding author: University of Pennsylvania, Philadelphia, PA 19104, USA; lx10077@upenn.edu.

Abstract

Watermarking provides a principled way to authenticate text generated by large language models. In practice, however, text may be mixed-source, with watermark evidence surviving at only a subset of token positions after rewriting, insertion, deletion, or paraphrasing. Although prior work has studied global watermark detection, when such signals can be localized remains unclear. We formulate watermark localization as a token-level multiple-testing problem based on pivotal statistics, with a latent indicator recording whether watermark dependence survives at each position. Under an asymptotic regime for signal sparsity, next-token concentration, and effective-vocabulary growth, we derive a sharp boundary for global detection and phase transitions for discovery and classification within the class of coordinatewise pivot-based localization rules. We show that discovery is strictly harder than detection and that consistent classification is impossible across the parameter regime within this class. We then develop an adaptive thresholding method that does not require knowledge of the exponents or time-varying next-token distributions, but uses a data-driven estimate of the surviving watermark fraction. The method attains the optimal discovery boundary and near-optimal discovery power relative to homogeneous pivot-based rules. Simulations support the theoretical phase transitions, while experiments on model-generated texts demonstrate practical localization performance under common edit mechanisms.

Key words: adaptive multiple testing; large language models; mixed-source text; pivotal statistics; statistical watermarking; watermark localization

1. Introduction

Large language models (LLMs) have recently become a powerful technology for generating human-like text and other media (Touvron et al., 2023a; Singh et al., 2025; Yang et al., 2025). They are now widely used in writing, education, programming, scientific discovery, and many other aspects of daily life. At the same time, their ability to generate fluent text at scale raises serious concerns about misuse, including misinformation (Weidinger et al., 2022; Starbird, 2019; Pan et al., 2023), academic integrity (Stokel-Walker, 2022; Milano et al., 2023), and data pollution for training future models (Radford et al., 2023; Shumailov et al., 2024). These risks make it increasingly important to reliably authenticate

the origin of text, especially when such information supports decisions about authorship attribution, academic assessment, and accountability (Wu et al., 2025).

To address this problem, watermarking has been proposed as a principled approach for verifying the origin of LLM-generated text (Kirchenbauer et al., 2023; Aaronson, 2023; Li et al., 2025a). Its main idea is to embed a hidden statistical signal into the generation process through controllable and recoverable pseudorandomness. Since LLMs generate text sequentially through sampling each token from next-token prediction (NTP) distributions, a watermarking scheme can privately modify this sampling rule so that the generated token still marginally follows the same NTP distribution, while being coupled with a pseudorandom variable known to the verifier. This coupling remains largely invisible to human users but provides valid statistical evidence for verifying whether the text was produced by a watermarked model. Following this principle, many watermarking schemes have been proposed since 2023 (Zhao et al., 2025a; Ji et al., 2026). These efforts have made watermarking one of the most promising approaches for providing provable evidence about the origin of LLM-generated text.

The practical relevance of text watermarking is no longer purely prospective. Google has deployed SynthID to watermark text generated by Gemini (Dathathri et al., 2024), and Anthropic has recently begun deploying embedded text watermarks in newly launched Claude models (Anthropic, 2026). These industry deployments highlight the growing practical importance of understanding whether watermark evidence remains reliable and informative after generated text is edited.

In real-world applications, however, LLM-generated text is rarely used without modification. Users often paraphrase, rewrite, insert, or delete portions of generated outputs before using or submitting them. Such edits weaken watermark signals because modified tokens may no longer preserve the original dependence between the text and the pseudorandomness, while the verifier observes only the final text. This challenge has motivated work on robust watermark design against human edits (Kuditipudi et al., 2024; Yoo et al., 2023; Zhu et al., 2024; Hou et al., 2024; Ren et al., 2024; Christ and Gunn, 2024; Moitra and Golowich, 2024). From a statistical perspective, Li et al. (2026) models edited text as a mixture in which only a fraction of tokens still carry watermark signals and studies global detection of partially watermarked text. Under the same mixture model, Li et al. (2025b) estimates how much watermark signal remains after common human editing. These studies address only aggregate questions: whether a mixed-source text still contains watermark evidence in aggregate, and how strong the remaining signal is.

A natural next question is more fine-grained: for a mixed-source text, can we identify which parts still preserve watermark evidence? We refer to this task as watermark localization, or watermark discovery, following the terminology of signal discovery in the multiple testing literature (Cai and Sun, 2017); see Figure 1 for an illustration. Unlike global detection, which only determines whether watermark evidence is present in the text as a whole, localization aims to provide token-level information about which positions still provide watermark evidence. This refinement is important when a document is only

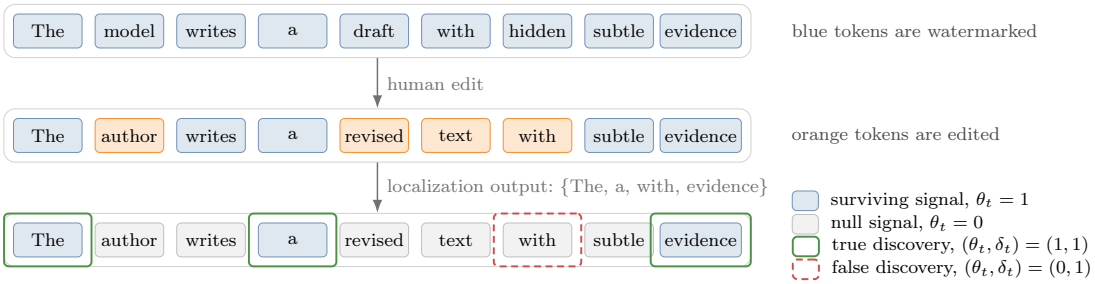

Fig. 1: Watermark localization under common human edits. For each final-text position t, $\theta_t = 1$ indicates a surviving watermark signal, and $\delta_t = 1$ indicates selection by the localization method. Selected positions with $\theta_t = 1$ are true discoveries, while those with $\theta_t = 0$ are false discoveries.

partially generated, collaboratively written, or substantially revised: a document-level conclusion may be too coarse to distinguish a short AI-generated passage from a largely AI-generated document, or to separate machine-generated content from substantial human revisions. Localization, therefore, provides a more informative basis for attribution, credit assignment, and targeted review.

Despite its practical importance, watermark localization remains much less understood from a statistical viewpoint. Recent works have explored this fine-grained problem using change-point detection (Li et al., 2024b) or online learning ideas (Zhao et al., 2025c), showing that localization is achievable in practice. However, these works are mainly algorithmic and do not characterize the statistical limits of localization for mixed-source data. This gap is nontrivial: mixed-source text creates a heterogeneous sequence in which watermark-preserving locations may be sparse, non-contiguous, and distributionally heterogeneous, while the watermark signal itself varies across positions due to autoregressive LLM generation. Motivated by this gap, we ask the following questions: *for mixed-source text, when is watermark localization statistically possible, and can it be achieved adaptively whenever localization is information-theoretically possible, without requiring prior knowledge of the source-mixing process?*

1.1. Our Contributions

A robust multiple-testing framework for watermark localization. We study these questions by developing a statistical theory and adaptive methodology for watermark localization in mixed-source LLM text. Our first contribution is a statistical formulation of this problem. For a text of length n, we use a scalar pivotal statistic Y_t from Li et al. (2025a) to quantify the watermark signal at each token position t. The key property is that, when no watermark signal survives at position t, Y_t follows a known null distribution μ_0, regardless of the marginal token distribution. This distribution-free null property has been central in prior statistical analyses of LLM watermarks and allows us to handle unknown and time-varying NTP distributions (Li et al., 2025a, 2026, 2025b).

To model the mixed-source text induced by human edits, we refine the mixture model of Li et al. (2026, 2025b) to the token level. For each position t, we introduce a latent survival indicator $\theta_t \in \{0,1\}$: $\theta_t = 1$ means that the watermark dependence at position t survives editing, while $\theta_t = 0$ means that this dependence is erased. Conditional on θ_t, the pivotal statistic follows either the null law μ_0 or a watermark-induced alternative law $\mu_{1,\boldsymbol{P}_t}$, determined by the local NTP distribution $\boldsymbol{P}_t$ of token w_t. Thus, given the observed sequence $Y_{1:n} := (Y_1, \ldots, Y_n)$, watermark localization can be viewed as the task of inferring the latent survival indicators $\theta_{1:n} := (\theta_1, \ldots, \theta_n)$.

Since exact recovery may be statistically impossible (Cai and Sun, 2017), we study this localization task through three inference goals of increasing strength. *Global detection* is the coarsest goal: it asks whether the verifier can reliably determine from the observed sequence whether at least one position satisfies $\theta_t = 1$. *Discovery* asks whether the verifier can identify a non-trivial set of watermark-preserving positions while ensuring a vanishing false discovery rate. Here, a false discovery is a selected position with no surviving watermark signal ($\theta_t = 0$), while a missed discovery is an unselected position with surviving watermark signal ($\theta_t = 1$). *Classification* is the strongest goal: it asks whether the verifier can asymptotically separate null positions from watermark-preserving positions across the entire text, with both false discoveries and missed discoveries vanishing. This hierarchy turns localization into a sequence of increasingly demanding inference goals. We next characterize, for each goal, when it is information-theoretically achievable.

Phase transitions for three inference goals. To characterize the fundamental limits of these three inference goals, we consider an asymptotic regime indexed by three exponents (p, q, α). The surviving watermark fraction satisfies $\varepsilon_n \asymp n^{-p}$, so that the text contains approximately $n\varepsilon_n \asymp n^{1-p}$ watermark-preserving positions; thus, larger p corresponds to sparser surviving signals. Each NTP distribution $\boldsymbol{P}_t := (P_{t,w})_w$ satisfies $1 - \max_{w \in \mathcal{W}} P_{t,w} \asymp n^{-q}$, so that larger q corresponds to a more concentrated next-token distribution and hence weaker token-level watermark evidence. Finally, the effective low-probability vocabulary tail grows at rate n^{α}, with larger α providing more rare-token opportunities for distinctive local evidence. These exponents characterize the statistical difficulty of the problem and are not required as inputs to our procedure. Under this regime, we derive an explicit boundary for global detection and sharp phase transitions for discovery and classification within the class of coordinatewise pivot-based localization rules.

- **Detection boundary.** We first show that global detection is possible if and only if

$$\max\{p + q,\ 2p + q - \alpha\} < 1, \tag{1}$$

up to boundary cases. This generalizes the fixed-vocabulary detection boundary of Li et al. (2026) to the growing-vocabulary regime. We also show that the truncated goodness-of-fit detection method of Li et al. (2026) continues to attain this enlarged boundary adaptively.

- **Discovery boundary.** For the main localization goal, we prove that discovery is possible if and only if

$$p < \alpha \quad \text{and} \quad p + q < 1, \tag{2}$$

  up to boundary cases. This region (2) is strictly smaller than the detectable region in (1), implying that discovery is fundamentally harder than global detection, because it requires locating a watermark-preserving position with vanishing false discovery rate. In particular, when $\alpha = 0$, discovery is impossible though global detection is still possible in some regimes.

- **Impossible classification.** We further prove that consistent classification is impossible throughout the entire (p, q, α) regime within the class of coordinatewise localization rules. The reason is that the two requirements of classification are not compatible with each other. To make missed discoveries vanish, one must select almost all watermark-preserving positions, including those with weak evidence; but doing so inevitably selects too many null positions, so the false discovery rate cannot vanish. In this way, discovery can still be possible because it only needs a few strong watermark-preserving positions, whereas classification requires reliable recovery of all of them.

Adaptive method and optimality. Our third contribution is an adaptive localization method, `SPOT` (**S**canning **P**ivots **O**ver **T**hresholds), presented in Algorithm 1, for watermark discovery. The method is designed to achieve discovery throughout the statistically discoverable region without prior knowledge of the problem-dependent quantities that determine the boundary, including p, q, α, and the time-varying NTP distributions, although it uses a data-driven estimate of the overall surviving watermark fraction (Li et al., 2025b).

At a high level, `SPOT` searches for positions whose pivotal statistics fall in an extreme upper tail relative to the null law μ_0. Thus, the main challenge is to choose a data-dependent threshold that is high enough to control false discoveries but not so high that useful watermark-preserving positions are lost. To do so, `SPOT` scans a grid of tail thresholds, estimates a surrogate false discovery level from empirical tail counts, and selects the largest discovery set whose estimated false discovery level is controlled. This tail-based calibration avoids estimating the heterogeneous alternative distributions, which is often impractical due to the inaccessibility of time-varying NTP distributions.

We establish two optimality guarantees. First, `SPOT` is boundary-optimal: it achieves discovery throughout the discoverable region characterized in (2). Second, it has near-optimal discovery power, recovering, up to a constant factor, as many watermark-preserving positions as the best rule in a natural class of homogeneous pivot-based procedures under the same false-discovery constraint. Figure 2 shows that `SPOT` achieves better watermark localization performance than the baseline method (Zhao et al., 2025c) under common edits such as random substitution, insertion, and deletion.

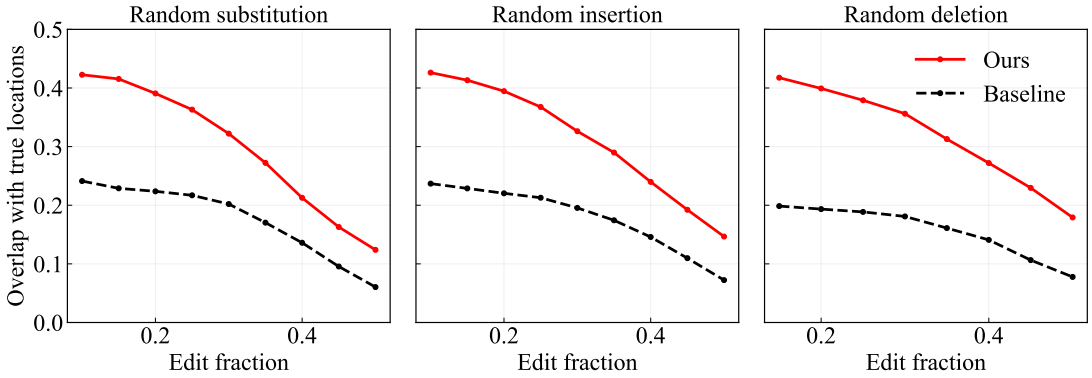

Fig. 2: LLM watermark localization under random edits; see Section 6 for details. For substitution, insertion, and deletion edits, we report the best intersection-over-union score subject to a false positive rate at most 0.05. Higher values indicate better localization performance.

1.2. Related Work

Watermark detection and localization. Our work is most closely related to statistical watermark detection and fine-grained localization. Li et al. (2026) studies global detection under a mixture model for edited text, and Li et al. (2025b) estimates the remaining watermark proportion in hybrid AI–human text. These works provide statistical tools for aggregate questions, but they do not identify where the surviving watermark signals are located. For localization, Li et al. (2024b) uses change-point detection to segment watermarked text, while Zhao et al. (2025c) proposes an online localization method for mixed-source text, which serves as our main empirical baseline. These methods show that localization is practically feasible, but they do not characterize its statistical limits or establish adaptive optimality. Our work addresses this gap by formulating watermark localization as a token-level multiple testing problem and deriving sharp limits for detection, discovery, and classification. More broadly, LLM watermarking has also been studied through biased (Kirchenbauer et al., 2023; Tsur et al., 2025) or distribution-preserving watermark design (Aaronson, 2023; Zhao et al., 2024; Xie et al., 2025; Giboulot and Furon, 2024; He et al., 2026b; Huang et al., 2026), robustness-oriented design (Kuditipudi et al., 2024; Yoo et al., 2023; Zhu et al., 2024; Hou et al., 2024; Ren et al., 2024; Christ and Gunn, 2024; Moitra and Golowich, 2024), and statistical detection (Li et al., 2025a, 2026; He et al., 2026a). Rather than proposing a new watermarking scheme, we focus on the statistical limits of localizing surviving watermark evidence in mixed-source text.

Sparse signal discovery and multiple testing. Our formulation is most directly connected to the signal discovery framework of Cai and Sun (2017), which studies how to identify sparse signals under false discovery control. We follow this perspective by distinguishing detection, discovery, and classification, but our setting differs because the alternative laws are heterogeneous and depend on unknown, time-varying NTP distributions generated by

an autoregressive LLM. Our work is also related to oracle compound decision rules and local-FDR methods for large-scale multiple testing (Sun and Cai, 2007), where local posterior information provides a natural principle for selecting non-null positions. In our setting, however, direct local-FDR estimation is difficult because the alternative law at each position is unobserved and varies with the local NTP distribution. We therefore use tail counts of pivotal statistics to control false discoveries without estimating the full heterogeneous alternative distribution. More broadly, our phase-transition analysis is related to sparse mixture detection and higher criticism, where sharp detection boundaries are characterized for rare and weak signals (Donoho and Jin, 2004, 2015), and to goodness-of-fit tests based on divergence statistics (Jager and Wellner, 2007).

1.3. Organization of the Paper

The remainder of the paper is organized as follows. In Section 2, we review the basics of LLM watermarking. In Section 3, we formulate the watermark localization problem and present our adaptive localization method. In Section 4, we investigate the information-theoretic limits of the problem and establish the optimality of our method. In Section 5, we present simulation studies to validate our theoretical findings. In Section 6, we conduct experiments on LLM-generated text to evaluate the empirical performance of our method. We conclude in Section 7 with a discussion of future research directions. All technical proofs and additional experimental details are provided in the Supplementary Material.

2. Preliminaries

Watermarking protocol. We study a standard watermarking protocol involving three parties: the model provider, the user, and the verifier (Kuditipudi et al., 2024; Xie et al., 2025; Li et al., 2026). As a motivating example, a student may use an LLM to assist with a homework assignment. The model provider embeds a statistical watermark into the generated text using a secret key `Key`, which is available to the verifier, such as an instructor, but hidden from the user. Before submission, the user may revise or edit the generated output. The verifier observes only the final text, which may already have been modified, and aims to localize the watermark signal, namely, to identify which parts of the text still preserve watermark signals. The verifier does not observe the original prompt, the unedited model output, or the model's internal parameters. Thus, the protocol consists of three stages: watermark embedding by the model provider, possible human editing by the user, and watermark localization from the final edited text by the verifier.

Watermark embedding. We next describe how a watermark is embedded during text generation. LLMs generate text autoregressively: at position t, given previous tokens $w_{1:(t-1)} := w_1 \cdots w_{t-1}$, the model computes a next-token prediction (NTP) distribution $\boldsymbol{P}_t := (P_{t,w})_{w\in\mathcal{W}}$ over the vocabulary $\mathcal{W}$ and samples the next token accordingly. A watermark modifies this sampling step in a secret-key-dependent way. Specifically, the model computes a pseudorandom variable $\zeta_t = \mathcal{A}(w_{(t-m):(t-1)}, \texttt{Key})$, where $\mathcal{A}$ is a

cryptographic hash function, Key is the watermarking key, and m is the context window size, and then selects the next token by a deterministic decoder $w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t)$. This creates a dependence between the generated token w_t and the pseudorandom variable ζ_t. The variable ζ_t is called pseudorandom because it behaves statistically like a random draw, while being exactly reproducible from $\mathcal{A}$, Key, and the local text $w_{(t-m):(t-1)}$. For theoretical analysis, we model ζ_t as i.i.d. samples from a reference distribution π, reflecting standard cryptographic idealizations (Barak, 2021; Schneier, 1996; Wu et al., 2024; Zhao et al., 2025b). A decoder $\mathcal{S}$ is unbiased if $\mathbb{P}_{\zeta\sim\pi}(\mathcal{S}(\boldsymbol{P}, \zeta) = w) = P_w$ for every NTP distribution $\boldsymbol{P} = (P_w)_{w\in\mathcal{W}}$ and candidate token w. Thus, unbiased watermarking preserves the marginal distribution of generated tokens while creating a hidden dependence on ζ_t that can later be used for verification.

Pivotal statistics for watermark evidence. Since our primary interest is watermark localization, we adopt the verifier's perspective. Given the final text $w_{1:n}$, the verifier reconstructs the corresponding pseudorandom sequence $\zeta_{1:n}$ using the hash function $\mathcal{A}$ and the shared key Key. Hence, the data for statistical analysis are the paired observations $\{(w_t, \zeta_t)\}_{t=1}^n$. Watermark evidence is encoded in the dependence between a token and its pseudorandom variable. If a token is human-written, or if its watermark dependence has been destroyed by editing, then w_t is independent of ζ_t; if the watermark signal is preserved, then w_t remains coupled with ζ_t through the decoder. To measure this dependence, we use a pivotal statistic $Y_t = Y(w_t, \zeta_t)$ (Li et al., 2025a), which is designed to follow a known null distribution μ_0 whenever w_t is independent of ζ_t, regardless of the marginal distribution of w_t. When the watermark dependence is preserved, Y_t instead follows an alternative distribution $\mu_{1,\boldsymbol{P}_t}$ determined by the local NTP distribution $\boldsymbol{P}_t$. Thus, pivotal statistics reduce watermark localization to distinguishing null-like positions from watermark-preserving positions. We formalize the resulting multiple testing problem in Section 3.1.

The Gumbel-max watermark. For concreteness, our theoretical analysis focuses on the Gumbel-max watermark (Aaronson, 2023), one of the most influential unbiased watermarking schemes. We use it as our main example because both its decoder and pivotal statistic admit explicit distributional forms, making it a clean setting for deriving sharp localization limits.[1] The Gumbel-max watermark is based on the Gumbel-max trick for sampling from a multinomial distribution (Gumbel, 1948; Papandreou and Yuille, 2011; Maddison et al., 2014; Jang et al., 2017). Let $\zeta = (U_w)_{w\in\mathcal{W}}$ consist of i.i.d. $U(0,1)$ random variables. The trick states that $\arg\max_{w\in\mathcal{W}} \log U_w/P_w$ follows the categorical distribution $\boldsymbol{P} = (P_w)_{w\in\mathcal{W}}$ exactly. This motivates the unbiased decoder

$$\mathcal{S}^{\mathrm{gum}}(\boldsymbol{P}, \zeta) := \arg\max_{w\in\mathcal{W}} \frac{\log U_w}{P_w}. \tag{3}$$

[1] These ideas are not specific to Gumbel-max and may extend to other schemes with suitable pivotal statistics.

For this watermark, the pivotal statistic is $Y_t = Y(w_t, \zeta_t) = U_{t,w_t}$, where $\zeta_t = (U_{t,w})_{w\in\mathcal{W}}$ collects the uniform pseudorandom at position t. If the observed token w_t is independent of ζ_t, then $Y_t \sim U(0,1)$. If the token is generated via the Gumbel-max decoder $\mathcal{S}^{\text{gum}}$, larger pseudorandom values are more likely to be selected, and the pivot becomes stochastically larger. More precisely, for a given NTP distribution $\boldsymbol{P}$, its distribution is $\mu_{1,\boldsymbol{P}}(Y \le r) = \sum_{w\in\mathcal{W}} P_w r^{1/P_w}$ for $r \in [0,1]$ (Li et al., 2025a). This explicit null-versus-alternative structure is the basis for our localization analysis.

3. Method

3.1. Problem Formulation

We first describe the statistical data structure behind watermark localization. From Section 2, the verifier observes the final text $w_{1:n}$, reconstructs the pseudorandom variables $\zeta_{1:n}$, and computes the pivotal statistic $Y_t := Y(w_t, \zeta_t)$ at each position t. This scalar statistic measures the watermark evidence in the pair (w_t, ζ_t). Its key property is that, if the token is human-written or if human editing has broken its dependence on the pseudorandom variable, then w_t is independent of ζ_t, and Y_t follows a known null distribution μ_0 regardless of the marginal distribution of w_t. In contrast, if the watermark signal survives, then w_t remains coupled with ζ_t through the decoder, and Y_t follows an alternative distribution $\mu_{1,\boldsymbol{P}_t}$ determined by the local NTP distribution $\boldsymbol{P}_t$.

This null-versus-signal structure motivates a latent label $\theta_t \in \{0,1\}$, where $\theta_t = 1$ indicates that the watermark dependence at position t survives editing, and $\theta_t = 0$ indicates that it is erased. At the level of pivotal statistics, this structure is summarized as

$$Y_t \mid (\boldsymbol{P}_t, \theta_t) \sim (1-\theta_t)\mu_0 + \theta_t \mu_{1,\boldsymbol{P}_t} \quad \text{for} \quad t = 1, \dots, n. \tag{4}$$

The localization target is to identify the signal set $I := \{t : \theta_t = 1\}$, consisting of positions where watermark evidence remains. We also denote the non-watermarking set by $N := \{t : \theta_t = 0\}$. Therefore, watermark localization is a token-level multiple decision problem: at each position, the verifier decides whether the local state is null-like or watermark-preserving.

A *localization rule* is a binary sequence $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)$, where $\delta_t = 1$ means that position t is declared to preserve watermark evidence. For a given rule $\boldsymbol{\delta}$, we define $S_{\boldsymbol{\delta}} := \{t : \delta_t = 1\}$ as the associated *discovery set*, namely the set of positions selected by the rule $\boldsymbol{\delta}$ as watermark-preserving. In this paper, we focus on *local rules* that are separate on pivotal statistics at each position, denoted by $\mathcal{D}_n := \{\boldsymbol{\delta} : \delta_t = \phi_t(Y_t),\ t = 1, \dots, n\}$, where each $\phi_t : [0,1] \to \{0,1\}$ is a measurable function that may depend on the position t and the text length n. Unless otherwise stated, all decision rules belong to $\mathcal{D}_n$.

Remark 3.1 (Connection to prior formulation). Our formulation in (4) is motivated by the mixture model of Li et al. (2026), but differs in that we explicitly introduce the latent binary sequence $\{\theta_t\}_{t=1}^n$ to encode which token positions still preserve watermark

Algorithm 1 SPOT: **S**canning **P**ivots **O**ver **T**hresholds

Input: Given text $w_{1:n}$, hash function $\mathcal{A}$, secret key Key, pivot function Y, grid endpoints $0 < u_{\min} < u_{\max} < 1$, grid size M_n, target level $\lambda_n \in (0,1)$, and slack $\eta_n \in (0, \lambda_n)$.

1: **Compute pseudorandomness.** For each $t = 1, \ldots, n$, reconstruct $\zeta_t = \mathcal{A}(w_{(t-m):(t-1)}, \texttt{Key})$.

2: **Compute pivotal statistics.** For each $t = 1, \ldots, n$, compute $Y_t = Y(w_t, \zeta_t)$.

3: **Construct tail grid.** Set $\mathcal{U}_n := \{u_j = u_{\min} + (j-1)\frac{u_{\max} - u_{\min}}{M_n - 1} : j = 1, \ldots, M_n\}$

4: **Estimate tail mass.** For each $u \in \mathcal{U}_n$, compute

$$\widehat{S}_n(u) := \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}\{F_0(Y_t) > \tau(u)\}. \tag{5}$$

5: **Estimate surviving fraction.** Obtain an estimate $\widehat{\varepsilon}_n$ of the surviving watermark fraction.

6: **Estimate null proportion in the tail.** For each $u \in \mathcal{U}_n$, set

$$\widehat{T}_n(u) := \frac{(1 - \widehat{\varepsilon}_n) S_0(\tau(u))}{\widehat{S}_n(u) \vee n^{-1}} = \frac{(1 - \widehat{\varepsilon}_n) n^{-u}}{\widehat{S}_n(u) \vee n^{-1}}. \tag{6}$$

7: **Select adaptive threshold.** Let

$$\widehat{u}_n := \inf\Big\{ u \in \mathcal{U}_n : \ \widehat{T}_n(u) \le \lambda_n - \eta_n \Big\}, \qquad \widehat{\tau}_n := \tau(\widehat{u}_n) = 1 - n^{-\widehat{u}_n}. \tag{7}$$

8: **Finalize decision.** Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ by setting $\delta_t = 1$ if $Y_t > \widehat{\tau}_n$, and $\delta_t = 0$ otherwise.

Output discoveries. Return the estimated discovery set $S_{\boldsymbol{\delta}} = \{t \in \{1, \ldots, n\} : \delta_t = 1\}$.

dependence after human edits. This additional structure turns robust watermark detection into a localization problem.

3.2. Our Method: SPOT

We now present our method SPOT in Algorithm 1. At a high level, the method searches for positions whose pivotal statistics look unusually abnormal under the null law μ_0, and declares such positions as discoveries. Under the null case, $\theta_t = 0$, so that the p-value $p_t := 1 - F_0(Y_t)$ is i.i.d. $U(0,1)$, where $F_0(y) := \mu_0(Y \le y)$ is the CDF of the null distribution μ_0. Hence, very small p-values, or equivalently very large values of $F_0(Y_t)$, provide evidence that the corresponding positions may preserve watermark signal. Therefore, the problem reduces to choosing a threshold: positions with $F_0(Y_t)$ above this threshold are declared as discoveries.

The key difficulty is that the optimal threshold is hard to find. We want to output a discovery set $S_{\boldsymbol{\delta}}$ while controlling false discoveries. Here, a false discovery is a selected position whose watermark dependence has been erased, that is, a position with $\theta_t = 0$

but $\delta_t = 1$. A threshold that is too low may include too many false discoveries, while a threshold that is too high may remove many true discoveries. We introduce λ_n as the target level for controlling the fraction of false discoveries among the selected positions.

Ideally, the threshold should depend on problem-specific quantities, such as the surviving watermark fraction ε_n and the NTP distribution $\boldsymbol{P}_t$, which are unfortunately unavailable in practice. To address this issue, `SPOT` scans thresholds of the form $\tau(u) = 1 - n^{-u}$. For each candidate u, it computes the empirical tail mass $\widehat{S}_n(u)$, which is the fraction of positions with $F_0(Y_t) > \tau(u)$. Under the null case where $\theta_t = 0$, this tail probability is $S_0(\tau(u)) = 1 - \tau(u) = n^{-u}$. If the surviving watermark fraction is ε_n, then roughly a fraction $1 - \varepsilon_n$ of positions are null-like, so the expected null contribution to this tail is about $(1-\varepsilon_n)S_0(\tau(u))$. Replacing ε_n by an estimate $\widehat{\varepsilon}_n$ and comparing this estimated null contribution with the observed tail mass give a quantity $\widehat{T}_n(u) := (1 - \widehat{\varepsilon}_n)S_0(\tau(u))/(\widehat{S}_n(u) \vee n^{-1})$.[2] We interpret $\widehat{T}_n(u)$ as an empirical proxy for the mFDR at the candidate threshold $\tau(u)$.

To control false discoveries, the algorithm selects the smallest grid point u such that $\widehat{T}_n(u) \leq \lambda_n - \eta_n$, where η_n is a slack term used to guard against random fluctuations. The statistic $\widehat{T}_n(u)$ estimates the false-discovery level of the tail selected by the threshold $\tau(u) = 1 - n^{-u}$. Thus, a large value of $\widehat{T}_n(u)$ indicates that the selected tail may still be largely explained by null positions and is therefore not reliable enough. Once $\widehat{T}_n(u)$ falls below $\lambda_n - \eta_n$, the tail is estimated to contain sufficiently few false discoveries, with the slack accounting for data variability. Since $\tau(u)$ increases with u, choosing the smallest admissible u yields the largest discovery set among the grid thresholds that pass the false-discovery check. This allows the method to retain as many candidate watermark-preserving positions as possible while controlling false discoveries.

Remark 3.2 (Choice of fraction estimator). Our theory is not tied to a specific estimator of the surviving watermark fraction. It only requires the accuracy condition in (9). In our implementation, we use the optimal estimator from Li et al. (2025b), which is also used in the experiments.

Remark 3.3 (Connection to prior methods). Our method `SPOT` is related to the signal discovery framework of Cai and Sun (2017), which studies Gaussian mixture models and uses local false discovery rate ideas through density estimation. The main challenge in our setting is that the data distributions in (4) are highly heterogeneous, since they depend on unknown and time-varying NTP distributions. Direct density estimation is thus difficult. Instead, `SPOT` uses tail probabilities $\widehat{S}_n(u)$ to estimate the false discovery level and selects the tail threshold adaptively for token-level localization.

4. Theoretical Guarantees

In this section, we establish the theoretical properties of our method. To facilitate the analysis, we first introduce the general assumptions in Section 4.1 and the performance

[2] Here $a \vee b := \max\{a, b\}$; the term $\vee n^{-1}$ in the denominator is used for numerical stability.

measures in Section 4.2. We then characterize the statistical limits and phase transitions of the resulting inference tasks in Section 4.3. Finally, in Section 4.4, we show that our method is adaptively optimal and is near-optimal in the number of discoveries under a fixed marginal false-discovery control.

4.1. General Assumptions

We first impose a probabilistic assumption on the watermark-generation and human-editing process. Each observed token w_t is modeled through a latent source mixture: after editing, each position either preserves the watermark dependence and is generated by the watermarked decoder, or the dependence is erased and the token behaves as an ordinary draw from the same NTP distribution. The latent variable θ_t records this post-edit survival state. This assumption provides the token-level basis for the null-versus-signal structure of the pivotal statistic Y_t.

Assumption 4.1 (Watermark generation and editing mechanism)**.** *Let $\boldsymbol{P}_t$ denote the NTP distribution at position t. Define the generation filtration $\mathcal{F}_{t-1} := \sigma(\{w_j, \zeta_j, \boldsymbol{P}_{j+1}\}_{j=1}^{t-1})$, which contains the history used to form $\boldsymbol{P}_t$. Let $\mathcal{H}_t := \sigma(\theta_1, \ldots, \theta_t)$ be the editing filtration and $\mathcal{G}_t := \mathcal{F}_{t-1} \vee \mathcal{H}_t$ be the joint generation-editing filtration. We assume the following.*

(a) ***Perfect pseudorandomness.*** *$\zeta_1, \ldots, \zeta_n$ are i.i.d., and ζ_t is independent of $\mathcal{G}_t$ for every t.*

(b) ***Latent source mixture.*** *Conditional on $\mathcal{G}_t$, the observed token is generated by*

$$w_t = \begin{cases} \mathcal{S}(\boldsymbol{P}_t, \zeta_t), & \theta_t = 1, \\ \text{an independent draw from } \boldsymbol{P}_t, & \theta_t = 0. \end{cases}$$

In the second case, w_t is conditionally independent of ζ_t.

(c) ***Surviving watermark fraction.*** *There exist constants $0 < c \le C < \infty$, independent of t and n, such that*

$$c \cdot \varepsilon_n \le \mathbb{P}(\theta_t = 1 \mid \mathcal{F}_{t-1}) \le C \cdot \varepsilon_n \quad \text{a.s. for all} \quad t \ge 1.$$

Assumption 4.1 separates the generation process from the editing process. Condition (a) is the standard idealization that the cryptographic pseudorandomness behaves as fresh randomness (Kirchenbauer et al., 2023; Li et al., 2025a) and is independent of the past and of the editing decision (Li et al., 2026, 2025b). Condition (b) encodes the local null-versus-signal structure with θ_t as a latent survival indicator (Li et al., 2026): when $\theta_t = 1$, the token w_t is produced by the watermarked decoder $\mathcal{S}$ and remains coupled with ζ_t; when $\theta_t = 0$, the token w_t has the same marginal NTP distribution but is independent of ζ_t, since human writing does not have access to the pseudorandom variable. The third row of Figure 1 illustrates this structure, where the ground-truth

vector is $(\theta_1, \ldots, \theta_9) = (1, 0, 0, 1, 0, 0, 0, 0, 1)$. Condition (c) controls the overall amount of surviving watermark evidence: ε_n is the surviving watermark fraction, up to constant factors, while the editing decision is allowed to depend on the previously generated text. Altogether, Conditions (b) and (c) characterize the mixture relation between the observed token w_t and the reconstructed pseudorandom variable ζ_t. Since the verifier does not know which positions still preserve the watermark dependence, the latent indicators θ_t provide a principled way to encode this uncertainty at the token level. Thus, the assumption does not model human editing behavior in detail, but preserves the essential statistical structure needed for analyzing watermark localization.

Assumption 4.2 (Asymptotic regime (p, q, α))**.** *For a text of length n, the surviving watermark fraction satisfies $\varepsilon_n \asymp n^{-p}$ for some $p \in [0, 1]$.*[3] *When $p = 0$, we further assume that ε_n is bounded away from one: there exists $\pi_\star \in (0, 1)$ such that $\varepsilon_n \le \pi_\star$ for all sufficiently large n. For each position t, the vocabulary $\mathcal{W}_n$ admits the decomposition*

$$\mathcal{W}_n = \{w_{t,n}^\star\} \cup L_{t,n} \cup C_{t,n}, \qquad P_{t,w_{t,n}^\star} = 1 - \Delta_n, \quad \Delta_n \asymp n^{-q}.$$

Here $w_{t,n}^\star$ is the dominant token under $\boldsymbol{P}_t$ and may vary with t. Thus, the dominant token can change across positions, but its probability is assumed to have the same asymptotic form. The residual mass is decomposed as

$$\Delta_n = \Delta_n^{\text{light}} + \Delta_n^{\text{core}}, \qquad \Delta_n^{\text{light}} \asymp n^{-q}, \qquad \Delta_n^{\text{core}} \le c\, n^{-q}.$$

The light set satisfies $|L_{t,n}| \asymp n^\alpha$ and $P_{t,w} \asymp n^{-(\alpha+q)}$ for all $w \in L_{t,n}$. The core set satisfies $|C_{t,n}| \asymp n^r$ for some $r < \alpha$, and there exist constants $0 < c_C \le C_C < \infty$ such that

$$c_C\, n^{-(\alpha+q)} \le P_{t,w} \le C_C\, n^{-(r+q)} \qquad \text{for all } w \in C_{t,n},\ t = 1, \ldots, n.$$

Consequently, the whole vocabulary satisfies $|\mathcal{W}_n| \asymp n^\alpha$. In the special case $q = 0$, we additionally assume that there exists $\eta_\star \in (0, 1)$ such that $P_{t,w_{t,n}^\star} \ge \eta_\star$ for all t and all sufficiently large n.

To study the statistical limits of watermark localization, we consider the asymptotic regime in Assumption 4.2, in the spirit of Li et al. (2025a, 2026). The regime is parameterized by (p, q, α). The parameter p describes how quickly the surviving watermark fraction $\varepsilon_n \asymp n^{-p}$ decays after human edits. The parameter q measures how concentrated each NTP distribution is around its dominant token: larger q means that the dominant token has probability closer to one, and hence the watermark signal becomes weaker. The parameter α describes the growth of the low-probability vocabulary tail.

[3] For two positive sequences a_n and b_n, we write $a_n \asymp b_n$ if there exist universal constants $0 < c_1 \le c_2 < \infty$, independent of n, such that $c_1 b_n \le a_n \le c_2 b_n$ for all sufficiently large n.

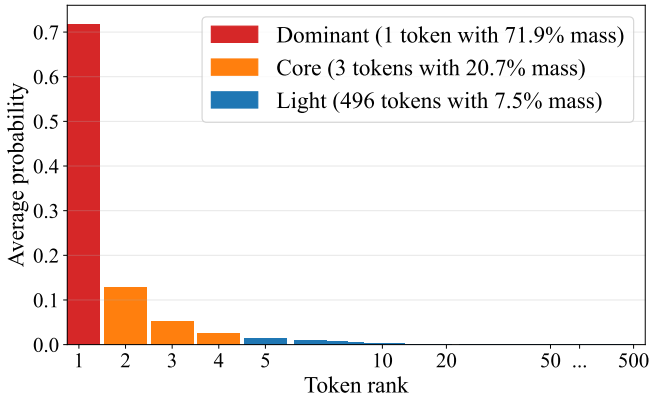

Fig. 3: Empirical illustration of the dominant–core–light structure in Assumption 4.2. We sample prompts from C4 datasets (Raffel et al., 2020) and generate continuations using OPT-1.3B (Zhang et al., 2022), then compute the average ranked next-token probabilities using temperature 0.5.

The dominant–core–light decomposition in Assumption 4.2 aims to capture a common shape of LLM next-token distributions. At each position t, the NTP distribution $\boldsymbol{P}_t$ may have a dominant prediction $w^\star_{t,n}$, and this dominant token is allowed to vary with the context. We only require its probability to follow the common scale $1-\Delta_n$, with $\Delta_n \asymp n^{-q}$. The remaining probability mass Δ_n is split into a core set $C_{t,n}$ and a light set $L_{t,n}$. The light set $L_{t,n}$ contains most of the vocabulary, with $|L_{t,n}| \asymp n^{\alpha}$, the same order as $|\mathcal{W}_n|$, and each light token has probability of order $n^{-(\alpha+q)}$. In contrast, the core set $C_{t,n}$ is much smaller, with $|C_{t,n}| \asymp n^{r}$ for some $r < \alpha$, but contains relatively larger-probability alternatives, with probabilities ranging from order $n^{-(\alpha+q)}$ to $n^{-(r+q)}$. Thus, the light set represents the large low-probability tail that drives vocabulary growth, while the core set represents a smaller group of more likely alternatives. This distinction explains why α controls the amount of tail information available for localization. The decomposition is also consistent with the empirical heavy-tailed behavior of language distributions (Zipf, 2016; Moreno-Sánchez et al., 2016), where a few tokens receive most of the probability mass while many weak alternatives remain available. Figure 3 provides an empirical illustration of this pattern using next-token distributions from OPT-1.3B on C4 prompts.

Remark 4.1 (Interpretation of α). The vocabulary of any deployed model is fixed. The growing-tail asymptotic should instead be viewed as a triangular-array approximation to successive model generations, which often adopt larger token vocabularies; for example, the Llama vocabulary increased fourfold, from 32,000 tokens in Llama 2 to 128,000 in Llama 3 (Touvron et al., 2023b; Dubey et al., 2024). Thus, α represents growth in the effective rare-token support rather than literal vocabulary growth within a single text.

The parameters (p, q, α) are used only to characterize theoretical difficulty; our method in Section 3.2 does not require knowing them. Instead, they allow us to state sharp

phase-transition results: p controls the sparsity of surviving watermark signals, q controls token-level signal strength through NTP concentration, and α controls how much useful tail information is available for localization. The additional condition for $q = 0$ only prevents the dominant-token probability from vanishing; when $q > 0$, this is automatic for sufficiently large n.

4.2. Performance Measures

We now introduce the performance measures used in our theoretical analysis. For completeness, we first formally define global detection. We then focus on token-level localization: following the signal discovery framework of Cai and Sun (2017), we define false positive and missed discovery rates, and use them to formalize discovery and classification.

Definition 4.1 (Global detection)**.** *A sequence of tests $\psi_n = \psi_n(Y_{1:n}) \in \{0, 1\}$ achieves* detection *if the sum of Type I and Type II errors vanishes for the global testing problem*

$$H_0 : \ Y_t \sim \mu_0 \text{ for all } t \qquad \text{versus} \qquad H_1 : \ Y_t \mid \boldsymbol{P}_t \sim (1 - \varepsilon_n)\mu_0 + \varepsilon_n \mu_{1,\boldsymbol{P}_t} \text{ for all } t, \quad (8)$$

that is,

$$\mathbb{P}_{H_0}(\psi_n = 1) + \mathbb{P}_{H_1}(\psi_n = 0) \to 0.$$

Here $\psi_n = 1$ means that the text is declared to contain surviving watermark signal. We say the testing problem in (8) *is* detectable *if there exists a sequence of tests that achieves global detection.*

Throughout the localization definitions below, decision rules are understood to belong to the local class $\mathcal{D}_n$ introduced in Section 3.2, unless explicitly stated otherwise. Global detection is different: it is a document-level testing problem, and the test ψ_n may use the full pivot sequence $Y_{1:n}$.

Definition 4.2 (Counts of discoveries and missed signals)**.** *Given a rule $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ and ground-truth labels $\theta_t \in \{0, 1\}$, define*

$$\mathrm{FP}_{\boldsymbol{\delta}} = \sum_{t:\theta_t=0} \delta_t, \qquad \mathrm{TP}_{\boldsymbol{\delta}} = \sum_{t:\theta_t=1} \delta_t, \qquad \mathrm{FN}_{\boldsymbol{\delta}} = \sum_{t:\theta_t=1} (1 - \delta_t).$$

Here $\mathrm{FP}_{\boldsymbol{\delta}}$, $\mathrm{TP}_{\boldsymbol{\delta}}$, and $\mathrm{FN}_{\boldsymbol{\delta}}$ count false discoveries, true discoveries, and missed signals, respectively. Since these quantities are random, we define their expectations as

$$\mathrm{EFP}_{\boldsymbol{\delta}} = \mathbb{E}[\mathrm{FP}_{\boldsymbol{\delta}}], \qquad \mathrm{ETP}_{\boldsymbol{\delta}} = \mathbb{E}[\mathrm{TP}_{\boldsymbol{\delta}}], \qquad \mathrm{EFN}_{\boldsymbol{\delta}} = \mathbb{E}[\mathrm{FN}_{\boldsymbol{\delta}}].$$

Definition 4.3 (Marginal false discovery and missed discovery rates)**.** *With the above notation, define*

$$\mathrm{mFDR}_{\boldsymbol{\delta}} = \frac{\mathrm{EFP}_{\boldsymbol{\delta}}}{\mathrm{EFP}_{\boldsymbol{\delta}} + \mathrm{ETP}_{\boldsymbol{\delta}}}, \qquad \mathrm{MDR}_{\boldsymbol{\delta}} = \frac{\mathrm{EFN}_{\boldsymbol{\delta}}}{\mathrm{EFN}_{\boldsymbol{\delta}} + \mathrm{ETP}_{\boldsymbol{\delta}}}.$$

Here, mFDR_δ *is the ratio of the expected number of false discoveries to the expected total number of discoveries, while* MDR_δ *is the ratio of the expected number of missed signals to the expected total number of watermark-preserving positions.*

Definition 4.4 (Discovery)**.** *A sequence of decision rules* $\boldsymbol{\delta}$ *achieves* discovery *if*

$$\mathrm{mFDR}_{\boldsymbol{\delta}} \ \to \ 0 \qquad \textit{and} \qquad \mathbb{P}\left(|S_{\boldsymbol{\delta}}| \geq 1\right) \to 1.$$

Discovery requires at least one selected position while the marginal false discovery rate vanishes. The localization problem in (4) *is said to be* discoverable *if there exists a sequence of local decision rules achieving discovery.*

Remark 4.2 (Multiple discoveries). The discovery condition can be strengthened to $\mathbb{P}(|S_{\boldsymbol{\delta}}| \geq K) \to 1$ for any fixed integer $K \geq 1$. This modification does not affect the discovery boundary; see the proof of Theorem 4.2.

Definition 4.5 (Classification)**.** *A sequence of decision rules* $\boldsymbol{\delta}$ *achieves* classification *if*

$$\mathrm{mFDR}_{\boldsymbol{\delta}} \to 0 \qquad \textit{and} \qquad \mathrm{MDR}_{\boldsymbol{\delta}} \to 0.$$

Classification requires both false discoveries and missed watermark-preserving positions to vanish asymptotically. The localization problem in (4) *is said to be* classifiable *if there exists a sequence of local decision rules achieving classification.*

The above definitions form a hierarchy of inference goals. Global detection in Definition 4.1 is the coarsest task: it only asks whether surviving watermark signal exists somewhere in the text, without identifying any location. For localization, the marginal false discovery rate (mFDR) in Definition 4.3 measures how many selected positions are actually null, while the missed discovery rate (MDR) measures how many watermark-preserving positions are not selected. Based on these aspects, discovery in Definition 4.4 is the weakest successful localization goal: it requires finding at least one surviving watermark signal while keeping the mFDR vanishing. Classification in Definition 4.5 is stronger, requiring asymptotically correct separation of null and watermark-preserving positions. We use these criteria above to show that global detection can be possible in regimes where localization is not, and to characterize which levels of localization are statistically achievable.

4.3. Statistical Limits and Phase Transitions

Throughout this subsection, we work under Assumptions 4.1 and 4.2, so the difficulty of the problem is indexed by (p, q, α). We characterize the fundamental limits of the inference goals defined above by identifying, in this parameter space, when each goal is achievable or impossible. These achievable and impossible regions yield the phase transitions studied below.

Detection boundary. We begin with global detection in Definition 4.1, which only asks whether any surviving watermark signal is present in the text. Following the terminology in (Donoho and Jin, 2004, 2015; Li et al., 2026), we say that H_0 and H_1 in (8) *merge asymptotically* if the total variation distance between the joint distributions of $Y_{1:n}$ under H_0 and H_1 tends to zero. In this case, no test can reliably distinguish the two hypotheses. Conversely, if the two distributions separate asymptotically, detection is possible. The following theorem gives the detection boundary under the (p, q, α) regime.

Theorem 4.1 (Detection boundary)**.** *Under Assumptions 4.1— 4.2, let* $p, q \in [0, 1]$ *and* $\alpha \in [0, 1)$.

- *If* $\max\{p + q,\ 2p + q - \alpha\} > 1$*, then* H_0 *and* H_1 *merge asymptotically. Hence, for any test, the sum of Type I and Type II errors tends to* 1 *as* $n \to \infty$.
- *If* $\max\{p + q,\ 2p + q - \alpha\} < 1$*, then* H_0 *and* H_1 *separate asymptotically. The likelihood-ratio test that rejects* H_0 *when the log-likelihood ratio is positive has a vanishing sum of Type I and Type II errors.*

Theorem 4.1 gives the global detection boundary in the (p, q, α) regime. Away from the boundary case, detection is possible exactly when both $p + q < 1$ and $2p + q < 1 + \alpha$ hold. Equivalently, the active boundary is $p + q = 1$ when $p < \alpha$, and $2p + q = 1 + \alpha$ when $p \geq \alpha$. These two constraints reflect two ways in which global detection can fail. The condition $p + q < 1$ rules out the case where surviving watermark evidence is too sparse or too weak to produce visible tail deviations. The condition $2p + q < 1 + \alpha$ captures the aggregate contribution of the growing light tail: when α is larger, more low-probability alternatives are available, and their collective contribution can make H_1 distinguishable from H_0. This result also clarifies the role of vocabulary growth. When $\alpha = 0$, the boundary reduces to the fixed-vocabulary detection boundary of Li et al. (2026). When $\alpha > 0$, the detectable region expands because the light tail provides additional aggregate evidence. Thus, global detection can benefit from a growing vocabulary tail, although it still only answers whether surviving watermark signal exists somewhere in the text and does not identify its locations.

Discovery boundary. We then turn to discovery in Definition 4.4, the weakest form of token-level localization. Unlike global detection, discovery requires selecting actual token positions, and thus its analysis must control the dependence among local decisions. For

this purpose, we impose an additional weak-dependence condition to rule out pathological long-range dependence in the generation and editing process.

Assumption 4.3 (Geometric mixing)**.** *Let $\bar{\mathcal{G}}_t := \mathcal{F}_t \vee \mathcal{H}_t$ be the full generation-editing information up to time t, and define the future sigma-field $\bar{\mathcal{G}}^+_{t+k} := \sigma(w_s, \zeta_s, \boldsymbol{P}_{s+1}, \theta_s : s \geq t+k)$. Define*

$$\alpha_{\bar{\mathcal{G}}}(k) := \sup_{t \geq 0} \alpha(\bar{\mathcal{G}}_t, \bar{\mathcal{G}}^+_{t+k}) \quad \textit{for all} \quad k \geq 1,$$

where $\alpha(\mathcal{A}, \mathcal{B}) := \sup_{A \in \mathcal{A},\ B \in \mathcal{B}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|$ is the strong-mixing coefficient between two σ-fields. We assume that this dependence decays geometrically: there exist universal constants $C > 0$ and $\rho \in (0,1)$ such that $\alpha_{\bar{\mathcal{G}}}(k) \leq C\rho^k$ for all $k \geq 1$.

The coefficient $\alpha_{\bar{\mathcal{G}}}(k)$ measures the maximal dependence between the information up to time t and the information starting from time $t+k$; see, e.g., Bradley (2005). Thus, Assumption 4.3 requires the dependence between the past/current generation-editing process and the distant future to decay rapidly as the gap k grows. It does not impose exact independence. Rather, it rules out long-range dependence that would make token-level localization difficult to analyze.

This condition is compatible with how practical LLM-generated text is produced. Although transformers can attend to previous tokens within their context window (Vaswani et al., 2017), deployed autoregressive LLMs operate with finite context windows. From this perspective, an autoregressive LLM with a fixed context length can be approximately viewed as a finite-memory Markov process (Zekri et al., 2024). Empirical studies also suggest that long-context models do not use all positions in the context uniformly effectively: performance can degrade when relevant information is distant or poorly positioned in the prompt (Liu et al., 2024; Li et al., 2024a; An et al., 2025). Therefore, Assumption 4.3 allows both the generated tokens and the editing indicators to depend on previous text, but requires this dependence to weaken with distance. Technically, this weak-dependence condition provides the concentration control needed for empirical tail counts and false-discovery analysis in the discovery problem.

Theorem 4.2 (Discovery boundary)**.** *Under Assumptions 4.1–4.3, let $p, q \in [0,1]$ and $\alpha \in [0,1)$.*

- *If $p < \alpha$ and $p + q < 1$, then there exists a local decision rule that achieves discovery.*
- *If $p \geq \alpha$ or $p + q > 1$, then no local decision rule can achieve discovery.*

In particular, when the effective vocabulary size is fixed, that is, $\alpha = 0$, discovery is impossible.

Theorem 4.2 gives the phase transition for the weakest nontrivial localization task, discovery. The condition $p+q < 1$ ensures that the surviving watermark evidence is strong enough to produce sufficiently extreme token-level evidence. The additional condition

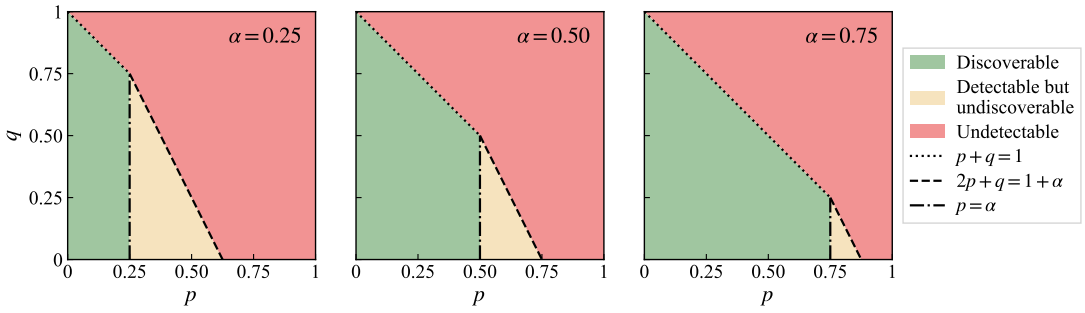

Fig. 4: Phase transitions for different values of $\alpha \in \{0.25, 0.50, 0.75\}$. The discoverable region is $p < \alpha$ and $p+q < 1$, while the undetectable region is $\max\{p+q,\ 2p+q-\alpha\} > 1$. The remaining region is detectable but not discoverable, illustrating the gap between detection and localization.

$p < \alpha$ is specific to localization: it requires the light set $L_{t,n}$ to grow fast enough relative to the sparsity of surviving watermark positions, so that at least one watermark-preserving position can stand out from null positions. This also explains why discovery is impossible when $\alpha = 0$: with a fixed effective vocabulary size, the light set does not grow, and the token-level separation needed for localization disappears.

This boundary shows that discovery is strictly harder than global detection. Detection can succeed by aggregating weak watermark evidence over the whole text, whereas discovery requires at least one individual position to be reliably identified. Consequently, there are regimes where the text can be detected as containing surviving watermark signal, but no local decision rule can reliably locate even one such position. In terms of phase boundaries, discovery requires both $p < \alpha$ and $p + q < 1$, while detection only requires $\max\{p + q, 2p + q - \alpha\} < 1$. Hence, away from boundary cases, the discoverable region is a strict subset of the detectable region. Figure 4 illustrates this separation. As α increases, the light set becomes larger and provides more low-probability alternatives, which strengthens token-level separation and makes both detection and discovery easier. At the same time, the detectable-but-undiscoverable region shrinks. In the degenerate case $\alpha = 0$, discovery is impossible, even though global detection may still be possible in some regimes.

Impossibility of classification. Discovery focuses on false-discovery control for the selected set, whereas classification additionally requires near-complete recovery of the signal set: almost all watermark-preserving positions must be selected while the selected set still contains only a negligible fraction of null positions. Under our general editing model, these two requirements cannot be made compatible. Indeed, $\theta_t = 1$ only means that the watermark dependence at position t survives; it does not guarantee that the pivotal statistic Y_t exhibits strong local evidence. As a result, some watermark-preserving positions may look nearly indistinguishable from null positions. To avoid missing these weak positions, a decision rule would have to select more aggressively, but doing so would also include too many null positions and prevent the false discovery rate from vanishing. Conversely, a

conservative rule may control false discoveries, but it must miss a non-negligible fraction of watermark-preserving positions. The next theorem shows that this incompatibility rules out classification throughout the (p, q, α) regime.

Theorem 4.3 (Impossibility of classification)**.** *Suppose Assumptions 4.1–4.2 hold. Then, for any $p, q \in [0, 1]$ and $\alpha \in [0, 1)$, no sequence of local decision rules can achieve classification.*

Remark 4.3 (Contrast with Gaussian mixtures). The impossibility in Theorem 4.3 contrasts with the Gaussian mixture setting of Cai and Sun (2017), where classification can be possible when the mean shift is sufficiently large. There, a stronger signal shifts the alternative distribution away from the null, so most non-null observations can be separated from null observations. In watermark localization, however, $\theta_t = 1$ only records the survival of watermark dependence; it does not imply that the pivotal statistic Y_t is far from the null law μ_0. Since watermark pivotal statistics are typically bounded, the null and alternative laws can still overlap substantially even when the watermark dependence survives. Thus, surviving watermark dependence is not analogous to a large mean shift, which explains why classification is impossible under the model considered here.

4.4. Adaptive Optimality

In this subsection, we establish the adaptive optimality of `SPOT` for discovery. Here, "adaptive" means that the algorithm does not require the problem-dependent parameters (Donoho and Jin, 2004, 2015), which in our setting include p, q, α and the NTP distributions $\boldsymbol{P}_{1:n}$. The term "optimal" refers to boundary optimality: the algorithm achieves discovery throughout the discoverable region.

Adaptive optimality for detection. We first revisit global detection for completeness. Although global detection is not the main focus of this paper, it is useful to ask whether an existing procedure can attain the detection boundary in the (p, q, α) regime. We show that this is the case for the robust detection method of Li et al. (2026), called `Tr-GoF`. This method is a truncated goodness-of-fit test: it compares the empirical distribution of the pivotal statistics with the null distribution μ_0 after a suitable truncation, and rejects the global null H_0 when the deviation is sufficiently large. Although `Tr-GoF` was originally developed for the fixed-vocabulary setting, the following theorem shows that it remains adaptively optimal under vocabulary growth.

Theorem 4.4 (Adaptive optimality for global detection)**.** *Suppose Assumptions 4.1–4.3 hold. If $\max\{p + q,\ 2p + q - \alpha\} < 1$, then* `Tr-GoF` *achieves global detection.*

Adaptive optimality for discovery. We next turn to the main focus of this paper: adaptive discovery. The key result is that `SPOT` attains the discovery boundary in Theorem 4.2 without knowing the parameters that determine this boundary. This adaptivity comes from

two ingredients. First, SPOT scans over M_n tail thresholds and selects the threshold from the data, rather than using the unknown optimal tail scale. Second, it uses an estimate $\widehat{\varepsilon}_n$ of the surviving watermark fraction, rather than requiring ε_n as prior knowledge. A feasible high-accuracy choice of $\widehat{\varepsilon}_n$ is the fraction estimator of Li et al. (2025b), which is also used in our experiments. For generality, the theorem below only requires the estimator to satisfy the accuracy condition in (9).

Theorem 4.5 (Adaptive optimality for discovery)**.** *Suppose Assumptions 4.1–4.3 hold. Run* SPOT *in Algorithm 1 with* $M_n \asymp (\log n)^a$ *for some* $a \geq 1$, $\lambda_n = C_1/\log n$, *and* $\eta_n = C_2/(\log n)^2$, *where* $C_1, C_2 > 0$ *are constants independent of* p, q, α, n. *Suppose the estimator* $\widehat{\varepsilon}_n$ *satisfies*

$$\mathbb{P}(|\widehat{\varepsilon}_n - \varepsilon_n| > c\eta_n) = o(n^{-1}) \tag{9}$$

for some constant $c > 0$. *If* $p < \alpha$ *and* $p + q < 1$, *then* SPOT *achieves discovery, provided that* $0 < u_{\min} < p + q < u_{\max} < 1$.

Remark 4.4 (Choice of the threshold grid). The condition $u_{\min} < p + q < u_{\max}$ is a coverage condition for the threshold grid. It ensures that the scan includes the relevant tail scale around $p + q$, where the adaptive threshold is selected asymptotically. This condition does not require prior knowledge of p or q: since discovery is possible only when $p + q < 1$, one may take $u_{\min}$ close to 0 and $u_{\max}$ close to 1 in practice to achieve it. In our experiments, we use $[u_{\min}, u_{\max}] = [0.005, 0.98]$ for simulations and $[0.05, 0.95]$ for LLM experiments.

Near-optimal discovery power. Beyond boundary optimality, we also ask whether SPOT discovers nearly as many watermark-preserving positions as possible under false positive control. To make this comparison meaningful and interpretable, we compare SPOT with homogeneous per-token local rules based on the pivotal statistic (Sun and Cai, 2007). Specifically, define

$$\mathcal{D}_n^{\mathrm{hom}}(\lambda_n) := \Big\{ \boldsymbol{\delta} : \ \exists\, \varphi : [0,1] \to \{0,1\} \text{ such that } \delta_t = \varphi(Y_t) \text{ for all } t, \ \mathrm{mFDR}_{\boldsymbol{\delta}} \leq \lambda_n \Big\}. \tag{10}$$

This class consists of rules that apply the same pivot-based decision function across positions and satisfy the same false positive rate constraint.

Theorem 4.6 (Near-optimal number of discoveries)**.** *Suppose the assumptions and setup in Theorem 4.5 hold. Let* $\boldsymbol{\delta}_{\texttt{SPOT}}$ *denote the decision rule returned by Algorithm 1 with the same parameters as in Theorem 4.5. Then there exist constants* $c_1, c_2 > 0$ *such that*

$$\mathrm{mFDR}_{\boldsymbol{\delta}_{\texttt{SPOT}}} \leq c_1 \lambda_n + o(1), \qquad \mathrm{ETP}_{\boldsymbol{\delta}_{\texttt{SPOT}}} \geq c_2 \cdot \sup_{\boldsymbol{\delta}' \in \mathcal{D}_n^{\mathrm{hom}}(\lambda_n)} \mathrm{ETP}_{\boldsymbol{\delta}'} + o(1).$$

Theorem 4.6 shows that SPOT is not only boundary-optimal but also quantitatively efficient. Among homogeneous local rules satisfying the target false positive rate constraint, SPOT achieves a constant fraction of the largest possible expected number of true discoveries, up to lower-order terms. Thus, the adaptive threshold chosen by SPOT does not only cross the correct phase boundary, but also retains near-oracle discovery power compared to a natural class of interpretable rules.

Remark 4.5 (The choice of benchmark class). We use $\mathcal{D}_n^{\mathrm{hom}}$ as the benchmark class because it matches the information available to a practical verifier. After observing the final text, the verifier can compute a pivotal statistic Y_t at each position, but does not observe the NTP distributions, the editing indicators, or other hidden generation states. Thus, a natural comparison is with rules that apply a common pivot-based decision function across positions under the same false positive rate constraint. We do not benchmark against fully time-varying or history-dependent rules, since such rules may rely on information unavailable to the verifier or on position-specific tuning that is not comparable to a practical localization procedure.

5. Simulations

In this section, we use simulations to verify the phase transition predicted by Theorem 4.2 and to illustrate the adaptive behavior of our SPOT.

5.1. Experimental Setup

We simulate pivotal statistics $Y_{1:n}$ from the mixture model in (4) to study the theoretical phase transitions. The simulation pipeline is summarized in Figure 5. We first specify the asymptotic parameters. For a given text length n and each triple (p, q, α), we set the surviving watermark fraction to $\varepsilon_n = 0.5n^{-p}$ and the residual mass away from the dominant token to $\Delta_n = n^{-q}$. In our implementation, we consider $\alpha \in \{0, 0.25, 0.5, 0.75\}$ and set $|\mathcal{W}_n| = 2 + \lfloor n^\alpha \rfloor$, with one core token and $\lfloor n^\alpha \rfloor$ light tokens. This corresponds to setting $r = 0$ in Assumption 4.2. Throughout this subsection, we use independent pseudorandom variables so that the simulation isolates the statistical structure of the localization problem.

Second, we generate each NTP distribution $\boldsymbol{P}_t$ using a latent Markov process $Z_t \in \{0, 1\}$. The role of Z_t is to allow the NTP distribution to vary mildly over time while keeping the same asymptotic structure. Specifically, Z_t is a two-state Markov chain started from stationarity, with $\mathbb{P}(Z_1 = 1) = 1/2$ and $\mathbb{P}(Z_t = Z_{t-1} \mid Z_{t-1}) = 0.95$. Thus, each regime, corresponding to a constant value of Z_t, tends to persist for many consecutive positions. For each regime $z \in \{0, 1\}$, we predefine an NTP distribution $\boldsymbol{P}_n^{(z)}$ satisfying the dominant–core–light structure in Assumption 4.2 and set $\boldsymbol{P}_t = \boldsymbol{P}_n^{(Z_t)}$.

Third, we generate the survival indicators $\{\theta_t\}$ independently of $\{Z_t\}$, but with temporal dependence across positions. The indicator $\theta_t \in \{0, 1\}$ records whether the watermark signal survives at position t. We sample $\{\theta_t\}$ from a two-state homogeneous Markov chain

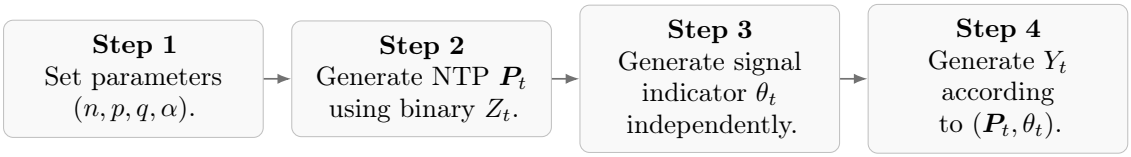

Fig. 5: Simulation pipeline for generating the pivotal statistics $\{Y_t\}_{t=1}^n$.

started from stationarity, with stationary mean $\mathbb{P}(\theta_t = 1) = \varepsilon_n$. Hence, on average, an ε_n fraction of positions retain watermark signal, typically in short bursts rather than in complete isolation.

Finally, given $(\boldsymbol{P}_t, \theta_t)$, we generate the pivotal statistic Y_t using the Gumbel-max watermark from Section 2. If $\theta_t = 0$, the watermark signal does not survive, and we sample Y_t independently from the null law μ_0. If $\theta_t = 1$, we sample directly from the watermark-induced alternative law $\mu_{1,\boldsymbol{P}_t}$ using the equivalent scalar construction $Y_t = U^{P_{t,w_t}}$, where $w_t \sim \boldsymbol{P}_t$ and $U \sim \mathrm{Unif}(0,1)$ are independent. Thus, each Y_t follows either the null μ_0 or the watermark-induced alternative law $\mu_{1,\boldsymbol{P}_t}$, according to θ_t, while the dependence across positions is inherited from the geometrically mixing process $(\boldsymbol{P}_t, \theta_t)$.

5.2. Discovery Boundary and Empirical Phase Transitions

We now visualize the empirical discovery boundary of `SPOT`. For a decision rule $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$, we evaluate its finite-sample discovery performance by the following *discovery error*

$$\mathrm{Err}_{\boldsymbol{\delta}} = \mathrm{mFDR}_{\boldsymbol{\delta}} + p_{\mathrm{miss}}(\boldsymbol{\delta}), \qquad p_{\mathrm{miss}}(\boldsymbol{\delta}) = \mathbb{P}(|S_{\boldsymbol{\delta}}| = 0), \tag{11}$$

where $S_{\boldsymbol{\delta}} = \{t : \delta_t = 1\}$ is the discovery set, $\mathrm{mFDR}_{\boldsymbol{\delta}}$ is the marginal false discovery rate, and $p_{\mathrm{miss}}(\boldsymbol{\delta})$ is the probability that the rule $\boldsymbol{\delta}$ makes no discovery. This error directly reflects the two requirements of discovery: controlling marginal false discoveries and ensuring a nontrivial discovery set (that contains at least one position). Smaller values therefore indicate better discovery performance.

In the implementation of `SPOT`, the target false discovery level is set as $\lambda_n = C/\log n$, where C is a calibration constant. For each parameter setting, we tune this calibration constant over a predetermined grid and report the smallest resulting error. This tuning is used only to visualize the empirical phase transition. In the following, we consider two complementary views: one-dimensional phase-transition slices and two-dimensional heatmap diagrams.

Phase transition for a fixed p or q. We first examine one-dimensional slices of the (p, q) plane. We either fix $q = 0.4$ and vary p, or fix $p = 0.6$ and vary q. For each $\alpha \in \{0, 0.25, 0.5, 0.75\}$ and $n \in \{10^2, 10^3, 10^4\}$, we run 200 independent Monte Carlo trials. For clarity, we present the representative slices for $\alpha \in \{0.5, 0.75\}$ here. Let $\mathcal{R}(a, b, K) = \{a + k(b-a)/(K-1) : k = 0, 1, \ldots, K-1\}$ denote K equally spaced grid points from a to b. For the fixed-q slices, we evaluate the error over $p \in \mathcal{R}(0.05, 0.95, 19)$; for the fixed-$p$ slices, we evaluate the error over $q \in \mathcal{R}(0.05, 0.95, 19)$. At each grid point, we tune the

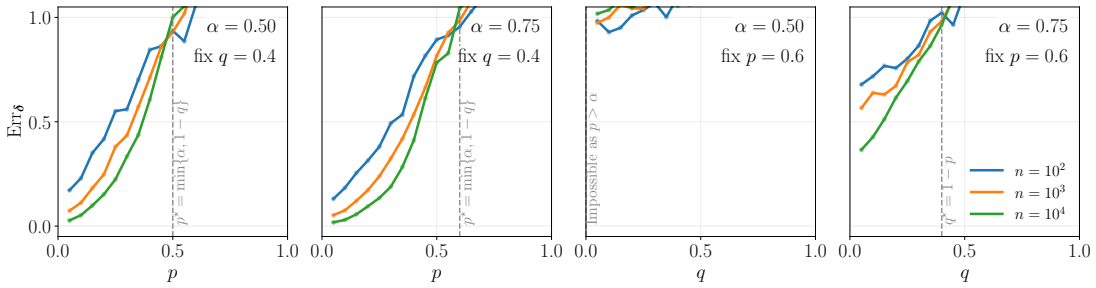

Fig. 6: Empirical discovery phase-transition slices for SPOT with $\alpha \in \{0.50, 0.75\}$. Each panel reports the smallest value of discovery errors defined in (11) over a calibration grid along one slice, averaged over 200 trials. The vertical dashed lines show the theoretical discovery boundary.

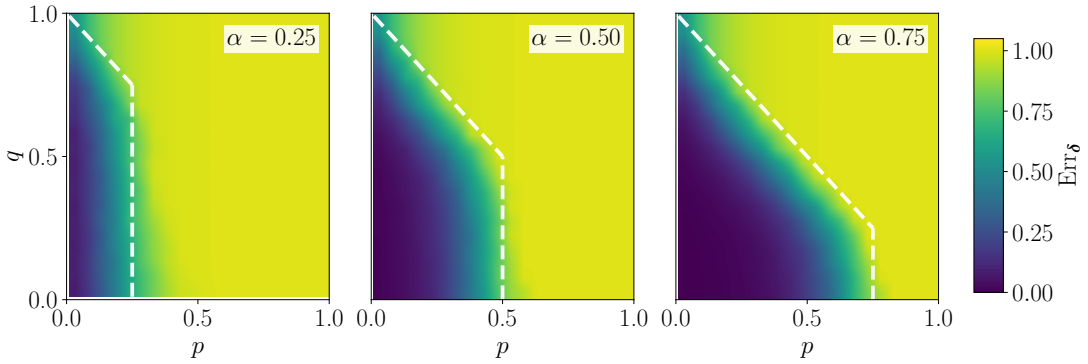

Fig. 7: Empirical discovery heatmap diagrams for SPOT with $n = 10^4$. Each panel reports the smallest value of discovery errors in (11) on a (p, q) grid, averaged over 500 trials, with displayed values clipped at 1.05. White dashed lines show the theoretical discovery boundaries.

calibration constant C over $\{0.05, 0.10, \ldots, 2.00\}$, set $\eta_n = 0$, and report the smallest error.

According to Theorem 4.2, the discoverable region is $\{(p, q) : p < \alpha,\ p + q < 1\}$. Thus, for fixed q, the critical transition in the p-direction is $p^\star(q; \alpha) = \min\{\alpha, 1 - q\}$. In particular, when $q = 0.4$, the predicted transition occurs at $p^\star(0.4; \alpha) = \min\{\alpha, 0.6\}$. For fixed p, discovery is possible only if $p < \alpha$, and then only for $q < 1 - p$. When $p = 0.6$, the predicted transition is therefore at $q = 0.4$ if $\alpha > 0.6$, while the entire slice is non-discoverable when $\alpha \leq 0.6$. Figure 6 confirms these predictions. For fixed $q = 0.4$, the error remains small when $p < p^\star(0.4; \alpha)$ and increases sharply after crossing the predicted boundary. For fixed $p = 0.6$, the transition occurs near $q = 0.4$ when $\alpha > 0.6$, while the error remains high across the whole slice when $\alpha \leq 0.6$. As n increases, the empirical transition becomes sharper and aligns more closely with the theoretical boundary.

Heatmap diagrams. We next evaluate the empirical boundary over a two-dimensional (p, q) grid. For the heatmap plots, we fix $n = 10^4$. For each $\alpha \in \{0.25, 0.5, 0.75\}$, we evaluate

the error over $p \in \mathcal{R}(0.01, 1, 20)$ and $q \in \mathcal{R}(\log_n(|\mathcal{W}_n|/(|\mathcal{W}_n|-1)), 1, 20)$.[4] For each (p, q), we run 500 independent Monte Carlo trials. At each grid point, we tune the calibration constant over 60 log-spaced values from 0.01 to 10, and report the smallest error. For visualization, displayed values are clipped at 1.05, so that all larger errors are shown as equally unfavorable.

Figure 7 shows the resulting heatmaps. Darker regions correspond to smaller discovery error, while lighter regions correspond to larger error. The dashed curves mark the theoretical boundaries $p = \alpha$ and $p + q = 1$ from Theorem 4.2. Across all $\alpha > 0$, the low-error region lies mostly inside the theoretically discoverable wedge $\{(p, q) : p < \alpha,\ p + q < 1\}$, while the high-error region dominates outside this wedge. Overall, the empirical transition bands align well with the predicted discovery boundary.

6. Open-source Model Experiments

6.1. Experiment Setup

We follow the setup of Li et al. (2026) and evaluate discovery performance on open-source LLMs under controlled edits. Specifically, we sample 1000 documents from the news-like subset of the C4 dataset (Raffel et al., 2020). For each document, we use the last 50 tokens as the prompt and ask OPT-1.3B (Zhang et al., 2022) to generate an additional $n = 400$ tokens as the continuation. During watermark generation, each pseudorandom variable is computed from the previous $m = 5$ tokens. We apply repeated-context masking, which adds the watermark only when the length-m prefix context has not appeared earlier in the generated history. This technique aims to reduce the frequency of watermarking and improve text quality (Dathathri et al., 2024). After obtaining the watermarked text, we apply the considered edit mechanisms to simulate human editing. We conduct experiments at temperatures $T \in \{0.3, 0.5, 0.7, 1\}$, ranging from low- to high-temperature generation.

Ground-truth labels and surviving fraction. For post-edit text, we define ground-truth labels by comparing the final text with the original pre-edit watermarked token sequence, following Li et al. (2025b). Specifically, after editing, we decode the edited text, re-tokenize it, and then pad or truncate it to the target length. For each position in the final post-edit sequence, we examine the block consisting of the current token and its previous m tokens. We label the current position as watermark-preserving only if this entire $(m + 1)$-token block appears contiguously in the original pre-edit sequence. We denote the resulting ground-truth labels by $\theta_1^{\mathrm{gt}}, \ldots, \theta_n^{\mathrm{gt}}$. The oracle surviving watermark fraction is then defined as $\varepsilon_{\mathrm{true}} = n^{-1} \sum_{t=1}^{n} \theta_t^{\mathrm{gt}}$.

Baseline methods. We compare `SPOT` with `AOL` (**A**daptive **O**nline **L**ocator), the token-level localization method in Algorithm 2 of Zhao et al. (2025c). Both methods output token-level decisions indicating whether each position preserves watermark evidence. For `SPOT`, we report two variants: `SPOT-oracle`, which uses the oracle surviving watermark fraction

[4] The lower bound on q ensures that the tail threshold $\tau(u) = 1 - n^{-u}$ is meaningful relative to the vocabulary size, since it enforces $1 - n^{-q} \geq 1/|\mathcal{W}_n|$.

$\varepsilon_{\text{true}}$, and `SPOT-plugin`, which uses the plug-in fraction estimator from Li et al. (2025b). Since all methods involve calibration constants, such as the constant C in $\lambda_n = C/\log n$ for `SPOT`, we report the best performance after tuning these constants over fixed grids.

Evaluation metrics. We consider several types of human modifications and evaluate token-level localization under each edit type. Let $S^\star$ denote the set of true watermark-preserving locations, and let $\widehat{S}$ denote the set of locations selected by a method. We report two localization metrics: the empirical true-positive rate (TPR), defined as $\widehat{\text{TPR}} = |\widehat{S} \cap S^\star|/|S^\star|$, and the intersection-over-union (IoU), defined as $\text{IoU} = |\widehat{S} \cap S^\star|/|\widehat{S} \cup S^\star|$. The former measures the fraction of true watermark-preserving locations recovered, while the latter provides a stricter overlap measure that penalizes both missed locations and extra selected locations. We also compute the empirical false-positive rate (FPR) as $\widehat{\text{FPR}} = |\widehat{S} \cap (S^\star)^c|/|(S^\star)^c|$, which measures the fraction of null locations incorrectly selected as watermark-preserving. While our theoretical discovery criterion is formulated in terms of mFDR, we use the conventional token-level FPR in the empirical comparison because it provides a common and directly interpretable operating point across localization methods. To compare methods under the same false-positive control, we report the largest empirical TPR and IoU each method can achieve subject to a prescribed empirical FPR upper bound.

6.2. Localization Performance

Following Li et al. (2026), we evaluate the localization performance of `SPOT` and `AOL` under three types of text modifications: (i) random edits, (ii) adversarial edits, and (iii) roundtrip translation. Random edits include substitution, insertion, and deletion. For a given edit rate, defined as the fraction of pre-edit tokens selected for modification, we randomly select that fraction of pre-edit tokens and either replace them, insert new tokens after them, or delete them. For substitutions and insertions, the new tokens are sampled uniformly from the vocabulary $\mathcal{W}$. Adversarial edits are more targeted: under the same edit-rate budget, they selectively modify pre-edit tokens to remove as much watermark signal as possible. Roundtrip translation translates the text from English to French and then back to English using another language model. Random and adversarial edits allow systematic control over the edit level, while roundtrip translation better reflects a practical text transformation.

Results for random edits. The random-edit results at temperature $T = 1$ are shown in Figure 8. Table 1 further reports the IoU and TPR averaged over those edit levels $\{0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$ for each temperature and each random edit type. As expected, increasing the edit fraction decreases IoU across all random edit types, since more watermark signals are removed or disrupted. In contrast, the TPR curves of `SPOT` are nearly flat as the edit fraction increases, showing that the method continues to recover a stable fraction of the surviving watermark-preserving locations under the same FPR constraint. This is consistent with the definition $\widehat{\text{TPR}} = |\widehat{S} \cap S^\star|/|S^\star|$: even though stronger

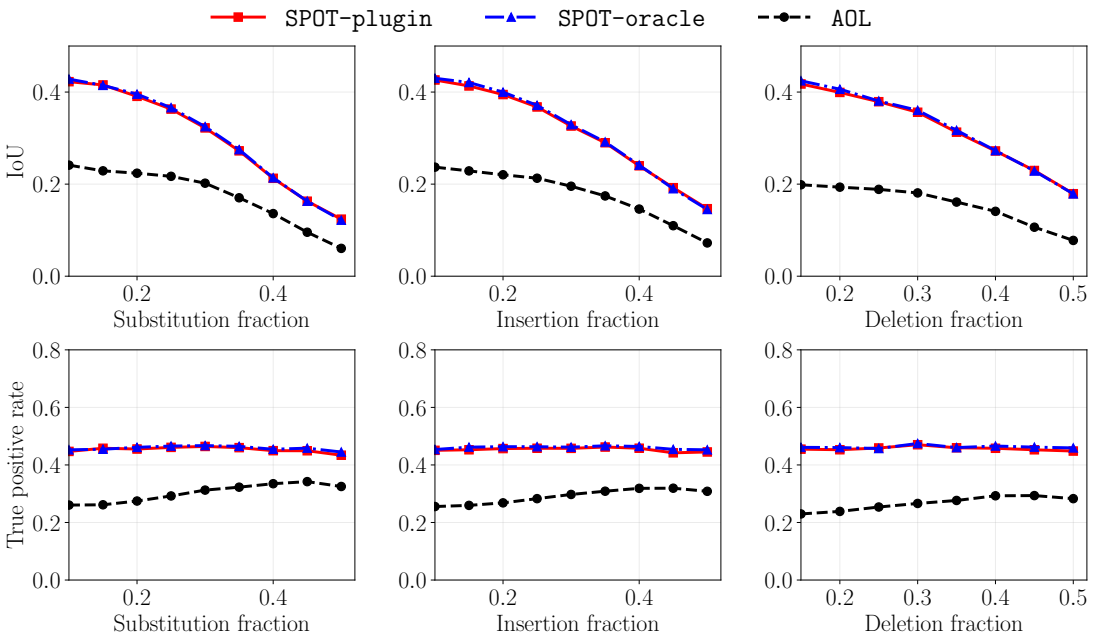

Fig. 8: Localization performance under three random edit mechanisms at temperature $T = 1$, measured by IoU (top row) and TPR (bottom row). Columns correspond to substitution, insertion, and deletion edits from left to right, with target FPR 0.05. Each panel reports the mean performance as a function of the edit fraction.

edits reduce the number of surviving locations, `SPOT` identifies a similar proportion of those that remain. Thus, the decline in IoU mainly reflects the increasing difficulty of overlap-based localization under heavier edits, while the stable TPR indicates that the tail-thresholding rule remains robust. We next examine how these patterns vary across methods and generation temperatures.

1. `SPOT` *performs better at moderate and high temperatures.* At $T = 1$, both `SPOT-oracle` and `SPOT-plugin` substantially outperform `AOL` across three random editings, in both IoU and TPR. The advantage is especially clear in the TPR curves in Figure 8: the TPR of `AOL` stays noticeably lower, while both variants of `SPOT` recover a much larger fraction of the surviving watermark-preserving locations across the whole range of edit fractions. At $T = 0.7$, the improvement becomes smaller but remains consistent. As shown in Table 1, `SPOT-oracle` remains better than `AOL` under most random-edit settings, and `SPOT-plugin` remains competitive. This suggests that `SPOT` is particularly effective when the LLM output is more diverse, where more usable watermark evidence remains available for localization.
2. `SPOT` *remains competitive at low temperatures in its oracle version.* At lower temperatures, the gap between methods becomes smaller because watermark localization is intrinsically harder. As shown in Table 1, at $T = 0.5$, `SPOT-oracle` remains comparable to `AOL`: it is better under random deletion and close under random substitution and insertion, both in IoU and TPR. By contrast, `SPOT-plugin` performs worse at this temperature,

suggesting that the degradation mainly comes from the reduced accuracy of the surviving-fraction estimator from Li et al. (2025b); we discuss this issue separately later. Finally, all methods have lower IoU and TPR at low temperatures. This is consistent with prior empirical observations that low-temperature generation makes LLM outputs more deterministic, leaves fewer usable watermark signals, and thereby makes watermark inference more difficult (Kirchenbauer et al., 2023; Lu et al., 2024; He et al., 2026a; Tsur et al., 2026).

Results for adversarial edits. We next consider adversarial edits, where the editor is assumed to have access to the hash function $\mathcal{A}$ and the secret key `Key` and can therefore target tokens carrying the strongest watermark signals. To approximate this setting, we first compute the pivotal statistics for the LLM-generated response, then identify the top-K tokens with the largest pivotal values, and finally replace them with randomly selected tokens. Here, K controls the edit budget and reflects the modification strength. Because these edits directly target the strongest watermark signals, they are more disruptive than random edits. The results are reported in the fourth row of Table 1.

The overall pattern is similar to that observed under random edits. At $T = 1$, both `SPOT-plugin` and `SPOT-oracle` outperform `AOL` by a clear margin in both metrics. At $T = 0.7$, `SPOT-oracle` and `AOL` have comparable performance, while `SPOT-plugin` remains competitive. At lower temperatures, the differences become smaller because the strongest watermark signals are already weak or removed. Overall, `SPOT-oracle` and `SPOT-plugin` show competitive or better performance than `AOL` in the adversarial setting.

Results for roundtrip translation. For roundtrip translation, the edit level is not directly controlled, so we compare methods under the same FPR constraints across temperatures. The IoU and TPR results are reported in the last row of Table 1. The qualitative pattern is broadly consistent with the random and adversarial edit settings. At $T = 1$, both `SPOT-oracle` and `SPOT-plugin` outperform `AOL` in both metrics. At moderate and low temperatures, `SPOT-oracle` remains comparable to `AOL`, while `SPOT-plugin` becomes less stable. This again suggests that the localization rule itself remains competitive, whereas the plug-in version can be limited by the quality of the surviving-fraction estimate.

Accuracy of the fraction estimator. The preceding experiments show a recurring gap between `SPOT-oracle` and `SPOT-plugin`, especially at lower temperatures and under stronger text modifications. To better understand this gap, we evaluate the accuracy of the fraction estimator $\widehat{\varepsilon}_n$ from Li et al. (2025b). Specifically, we report the relative root mean squared error

$$\frac{\mathrm{RMSE}(\widehat{\varepsilon}_n)}{\varepsilon_{\mathrm{true}}} = \frac{\sqrt{\mathbb{E}\left[(\widehat{\varepsilon}_n - \varepsilon_{\mathrm{true}})^2\right]}}{\varepsilon_{\mathrm{true}}}. \tag{12}$$

This quantity measures the typical estimation error on the scale of the true surviving fraction $\varepsilon_{\mathrm{true}}$. For example, values around 0.25, 0.5, and 1 correspond roughly to typical errors of 25%, 50%, and 100% of the true surviving fraction, respectively. Thus, values

Table 1. Average IoU and TPR across edit levels at different generation temperatures. For each metric, the tuning parameter is selected separately within the highest nonempty 0.01-wide empirical-FPR bin not exceeding 0.05. For random edits, entries are averaged over edit rates $\{0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$. For adversarial edits, entries are averaged over edit budgets $K \in \{10, 15, 20, 30, 40\}$. For roundtrip translation, the edit level is not directly controlled.

Edit Types	**Methods**	**IoU**				**TPR**			
		T = 1	**T = 0.7**	**T = 0.5**	**T = 0.3**	**T = 1**	**T = 0.7**	**T = 0.5**	**T = 0.3**
Random substitution	`AOL`	0.204	0.141	**0.086**	**0.039**	0.297	0.214	**0.154**	**0.090**
	`SPOT-plugin`	0.344	0.153	0.063	0.020	0.458	0.229	0.128	0.067
	`SPOT-oracle`	**0.347**	**0.160**	0.076	0.032	**0.461**	**0.239**	0.151	0.089
Random insertion	`AOL`	0.199	0.142	**0.089**	**0.044**	0.282	0.207	0.151	**0.092**
	`SPOT-plugin`	0.349	0.158	0.069	0.021	0.457	0.229	0.131	0.065
	`SPOT-oracle`	**0.354**	**0.165**	0.084	0.036	**0.461**	**0.237**	**0.155**	0.091
Random deletion	`AOL`	0.176	0.132	0.079	**0.039**	0.261	0.193	0.127	0.075
	`SPOT-plugin`	0.363	0.162	0.069	0.023	0.454	0.223	0.116	0.061
	`SPOT-oracle`	**0.370**	**0.175**	**0.085**	0.038	**0.462**	**0.239**	**0.137**	**0.078**
Adversarial edits	`AOL`	0.238	**0.140**	0.037	0.006	0.254	**0.162**	0.054	0.010
	`SPOT-plugin`	0.352	0.103	0.012	0.000	0.365	0.124	0.025	0.000
	`SPOT-oracle`	**0.376**	0.134	**0.050**	**0.008**	**0.389**	0.158	**0.092**	**0.026**
Roundtrip translation	`AOL`	0.217	0.150	**0.088**	**0.031**	0.301	0.213	**0.166**	0.081
	`SPOT-plugin`	0.344	0.153	0.055	0.014	0.426	0.208	0.102	0.032
	`SPOT-oracle`	**0.351**	**0.165**	0.081	0.028	**0.433**	**0.226**	0.162	**0.085**

Table 2. Relative root mean squared error defined in (12) of the Li-OPT fraction estimator at different temperatures. For random and adversarial edits, the relative RMSE is computed separately at each edit level and then averaged over the same edit levels as in Table 1. For roundtrip translation, the edit level is not directly controlled.

Edit Types	**T = 1**	**T = 0.7**	**T = 0.5**	**T = 0.3**
Random substitution	0.263	0.529	1.370	3.988
Random insertion	0.247	0.477	1.167	3.420
Random deletion	0.268	0.516	1.450	3.603
Adversarial edits	0.158	0.293	0.879	2.496
Roundtrip translation	0.289	0.432	1.397	4.115

substantially below one indicate that the plug-in estimate is reasonably accurate, whereas values near or above one indicate that fraction estimation can become a practical bottleneck.

Table 2 supports this interpretation. At high temperature, the relative RMSE is small for random and adversarial edits, and `SPOT-plugin` closely tracks `SPOT-oracle`. As the temperature decreases, the relative RMSE increases sharply, and the gap between the plug-in and oracle versions becomes larger. The effect is particularly visible at $T = 0.5$ and $T = 0.3$, where the relative RMSE is often close to or above one. These results suggest that the low-temperature degradation of `SPOT-plugin` is mainly driven by inaccurate estimation of the surviving watermark fraction, rather than by a failure of the localization rule itself.

Computational cost. Table 3 reports the wall-clock verification time at $T = 1$. The total cost consists of two components: recomputing the token-level pivotal statistics from the edited text and running the localization method once the pivots are available. Pivot recomputation is shared by all methods and dominates the overall runtime. Conditional on

Table 3. Average total verification time per sample at temperature $T = 1$, using the parameters selected to maximize IoU within the prescribed empirical-FPR. The total time is the sum of pivot recomputation time and method-inference time.

Edit case	`AOL`	`SPOT-plugin`	`SPOT-oracle`
Random substitution	0.232	0.221	0.216
Random insertion	0.229	0.218	0.213
Random deletion	0.117	0.114	0.108
Adversarial edits	0.229	0.219	0.213
Roundtrip translation	0.117	0.115	0.108
Average	0.185	0.177	0.172

the computed pivots, `SPOT-plugin` takes only 0.005–0.008 seconds per sample, compared with 0.009–0.016 seconds for `AOL`, showing that its improved localization performance incurs no additional computational overhead.

7. Discussion

This paper studies watermark localization in mixed-source LLM text from a statistical perspective. We formulate localization as a token-level multiple testing problem based on pivotal statistics, where each position has a latent indicator recording whether the watermark dependence survives editing. Under a regime that captures sparse surviving signals, concentrated next-token distributions, and growing vocabularies, we characterize the statistical limits of three inference goals: global detection, discovery, and classification. Our results show that discovery is strictly harder than global detection, and that consistent classification is impossible within the class of coordinatewise pivot-based localization rules. We then propose `SPOT`, an adaptive tail-thresholding method that does not require knowledge of the asymptotic exponents or the time-varying next-token distributions. The method achieves the optimal discovery boundary and attains near-optimal discovery power among natural local rules. Simulations support the theoretical phase transitions, while real-LLM experiments show the localization performance of `SPOT` under common edit mechanisms.

Several directions remain open. First, our current implementation uses the fraction estimator of Li et al. (2025b) to estimate the surviving watermark fraction. Although our theory allows any estimator satisfying a suitable accuracy condition, the experiments show that this step can become a practical bottleneck, especially at low temperatures where next-token distributions are more concentrated. This raises the question of whether localization can be performed without a separate fraction-estimation step. More broadly, the current procedure is batch-based: it uses a global fraction estimate for the entire text before choosing the localization threshold. An interesting direction is to develop online or streaming localization methods, where text arrives sequentially and the threshold is updated using accumulating evidence. Such methods may better adapt to documents whose source composition changes over time and reduce the need for a fixed global estimate of the surviving watermark fraction.

Second, our impossibility result shows that consistent classification is impossible within the class of coordinatewise localization rules. This leaves open the possibility of stronger recovery guarantees under additional structural assumptions or for localization procedures that exploit information across multiple token positions. Our model allows the latent survival indicators to switch frequently between watermark-preserving and null states, making exact token-level recovery too demanding. In practice, however, mixed-source text may have more block-like structure (Li et al., 2024b): LLM-generated passages, human-written passages, and heavily revised segments may each persist over multiple consecutive tokens. Under such segment-level regularity, or when the target is region-level rather than exact token-level recovery, stronger forms of localization may become possible.

Third, our analysis focuses on token-level watermark localization. For Gumbel-max and other token-level watermarking schemes, the extension is relatively direct whenever valid pivotal statistics can be constructed for individual token positions. Semantic or sentence-level watermarks are different (Hou et al., 2024; Ren et al., 2024; Huo et al., 2026): they may encode watermark information through sentence meanings, paraphrase-invariant features, or vector representations of larger text units rather than token-level dependence on pseudorandomness. In such settings, the localization unit may be a sentence, span, or semantic embedding, and the null-versus-signal formulation must be redefined. Extending localization theory to these non-token-level watermarks is an important open direction.

Data Availability

The C4 dataset used in this study is publicly available and cited in the manuscript. The code and materials required to reproduce all numerical results are included with the submission for review.

Acknowledgments

We thank Hossein Moradi Rekabdarkolaee for helpful comments on the exposition of an earlier version of this work during the NISS Writing Workshop. The authors used generative AI tools, including OpenAI Codex, to assist with language polishing and code development. All AI-assisted text and code were reviewed, edited, and validated by the authors, who take full responsibility for the manuscript and its results. This work was supported in part by NIH grants R01MH143267, U01CA274576, and R01EB036016, NSF grant DMS-2310679, a Meta Faculty Research Award, and Wharton AI for Business. J. Blanchet gratefully acknowledges support from the Department of Defense through ONR award 1398311 and from the National Science Foundation through grants 2312204 and 2403007. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

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

2. Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. Why does the effective context length of LLMs fall short? In *International Conference on Learning Representations*, volume 2025, pages 54791–54810, 2025.
3. Anthropic. How Claude marks AI-generated content, August 2026. URL `https://support.claude.com/en/articles/16266773-how-claude-marks-ai-generated-content`.
4. Boaz Barak. An intensive introduction to cryptography, lectures notes for Harvard CS 127. `https://intensecrypto.org/public/index.html`, Fall 2021.
5. Richard C Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144, 2005.
6. T. Tony Cai and Wenguang Sun. Optimal screening and discovery of sparse signals with applications to multistage high throughput studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):197–223, 2017.
7. Miranda Christ and Sam Gunn. Pseudorandom error-correcting codes. In *Annual International Cryptology Conference*, pages 325–347. Springer, 2024.
8. Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Merey, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, Ilia Shumailov, Ciprian Baetu, Sven Gowal, Demis Hassabis, and Pushmeet Kohli. Scalable watermarking for identifying large language model outputs. *Nature*, 634 (8035):818–823, 2024. doi: 10.1038/s41586-024-08025-4.
9. David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.
10. David Donoho and Jiashun Jin. Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical science*, 30(1):1–25, 2015.
11. Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
12. Eva Giboulot and Teddy Furon. WaterMax: Breaking the LLM watermark detectability-robustness-quality trade-off. *Advances in Neural Information Processing Systems*, 37:18848–18881, 2024.
13. Emil Julius Gumbel. *Statistical theory of extreme values and some practical applications: A series of lectures*, volume 33. US Government Printing Office, 1948.
14. Weiqing He, Xiang Li, Tianqi Shang, Li Shen, Weijie Su, and Qi Long. On the empirical power of goodness-of-fit tests in watermark detection. In *Advances in neural information processing systems*, volume 38, pages 18761–18793, 2026a.
15. Weiqing He, Xiang Li, Li Shen, Weijie J Su, and Qi Long. Improving the trade-off between watermark strength and speculative sampling efficiency for language models. In *International Conference on Learning Representations*, 2026b. URL `https://openreview.net/forum?id=HA8vzzT6Ax`.
16. Abe Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. SemStamp: A semantic watermark with paraphrastic robustness for text generation. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4067–4082, 2024.
17. Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=uWVC5FVidc`.
18. Baihe Huang, Eric Xu, Kannan Ramchandran, Jiantao Jiao, and Michael I Jordan. Towards anytime-valid statistical watermarking. *arXiv preprint arXiv:2602.17608*, 2026.
19. Jiahao Huo, Shuliang Liu, Bin Wang, Junyan Zhang, Yibo Yan, Aiwei Liu, Xuming Hu, and Mingxun Zhou. PMark: Towards robust and distortion-free semantic-level watermarking with channel constraints. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=EhDgP69DJG`.
20. Leah Jager and Jon A Wellner. Goodness-of-fit tests via phi-divergences. *Annals of Statistics*, 35(5):2018–2053, 2007.
21. Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=rkE3y85ee`.
22. Wenlong Ji, Weizhe Yuan, Emily Getzen, Kyunghyun Cho, Michael I Jordan, Song Mei, Jason Weston, Weijie J Su, Jing Xu, and Linjun Zhang. An overview of large language models for statisticians. *The American Statistician*, (just-accepted):1–106, 2026.
23. John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, volume 202, pages 17061–17084, 2023.
24. Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=FpaCL1MO2C`.

25. Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. Loogle: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333, 2024a.
26. Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *The Annals of Statistics*, 53(1):322–351, 2025a.
27. Xiang Li, Garrett Wen, Weiqing He, Jiayuan Wu, Qi Long, and Weijie J. Su. Optimal estimation of watermark proportions in hybrid AI–Human texts. *arXiv preprint arXiv:2506.22343*, 2025b. arXiv preprint.
28. Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J. Su. Robust detection of watermarks for large language models under human edits. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(2):491–515, 2026. doi: 10.1093/jrsssb/qkaf056. URL `https://doi.org/10.1093/jrsssb/qkaf056`.
29. Xingchi Li, Guanxun Li, and Xianyang Zhang. Segmenting watermarked texts from language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 14634–14665, 2024b.
30. Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.
31. Yijian Lu, Aiwei Liu, Dianzhi Yu, Jingjing Li, and Irwin King. An entropy-based text watermarking detection method. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11724–11735, 2024.
32. Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. In *Advances in Neural Information Processing Systems*, volume 27, 2014.
33. Florence Merlevède, Magda Peligrad, and Emmanuel Rio. Bernstein inequality and moderate deviations under strong mixing conditions. In Christian Houdré, Vladimir Koltchinskii, David M. Mason, and Magda Peligrad, editors, *High Dimensional Probability V: The Luminy Volume*, volume 5 of *IMS Collections*, pages 273–292. Institute of Mathematical Statistics, 2009. doi: 10.1214/09-IMSCOLL518.
34. Silvia Milano, Joshua A McGrane, and Sabina Leonelli. Large language models challenge the future of higher education. *Nature Machine Intelligence*, 5(4):333–334, 2023.
35. Ankur Moitra and Noah Golowich. Edit distance robust watermarks for language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 20645–20693, 2024.
36. Isabel Moreno-Sánchez, Francesc Font-Clos, and Álvaro Corral. Large-scale analysis of Zipf's law in english texts. *PloS one*, 11(1):e0147073, 2016.
37. Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. On the risk of misinformation pollution with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
38. George Papandreou and Alan L Yuille. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *International Conference on Computer Vision*, pages 193–200. IEEE, 2011.
39. Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.
40. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
41. Jie Ren, Han Xu, Yiding Liu, Yingqian Cui, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. A robust semantics-based watermark for large language model against paraphrasing. In *Findings of the Association for Computational Linguistics*, pages 613–625, 2024.
42. Bruce Schneier. *Applied Cryptography*. John Wiley & Sons, 1996.
43. Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.
44. Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.
45. Kate Starbird. Disinformation's spread: Bots, trolls and all of us. *Nature*, 571(7766):449–450, 2019.
46. C Stokel-Walker. AI bot ChatGPT writes smart essays—Should professors worry? *Nature News*, 2022.
47. Wenguang Sun and T Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, 2007.
48. Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.
49. Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv*

preprint arXiv:2307.09288, 2023b.

50. Dor Tsur, Carol Xuan Long, Claudio Mayrink Verdun, Sajani Vithana, Hsiang Hsu, Chun-Fu Chen, Haim H. Permuter, and Flavio Calmon. HeavyWater and SimplexWater: Distortion-free LLM watermarks for low-entropy distributions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=R5EBtNE2Y9.
51. Dor Tsur, Carol Xuan Long, Claudio Mayrink Verdun, Sajani Vithana, Hsiang Hsu, Chun-Fu Chen, Haim H. Permuter, and Flavio Calmon. HeavyWater and SimplexWater: Distortion-free LLM watermarks for low-entropy distributions. In *Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=R5EBtNE2Y9.
52. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
53. Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 214–229, 2022.
54. Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338, 2025.
55. Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. A resilient and accessible distribution-preserving watermark for large language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=c8qWiNiqRY.
56. Yangxinyu Xie, Xiang Li, Tanwi Mallick, Weijie Su, and Ruixun Zhang. Debiasing watermarks for large language models via maximal coupling. *Journal of the American Statistical Association*, 120(551):1424–1436, 2025.
57. An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
58. KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. Robust multi-bit natural language watermarking through invariant features. In *Annual Meeting Of The Association For Computational Linguistics*, 2023.
59. Oussama Zekri, Ambroise Odonnat, Abdelhakim Benechehab, Linus Bleistein, Nicolas Boullé, and Ievgen Redko. Large language models as Markov chains. *arXiv preprint arXiv:2410.02724*, 2024.
60. Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
61. Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for AI-generated text. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=SsmT8aO45L.
62. Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, et al. Sok: Watermarking for AI-generated content. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 2621–2639. IEEE, 2025a.
63. Xuandong Zhao, Lei Li, and Yu-Xiang Wang. Permute-and-Flip: An optimally stable and watermarkable decoder for LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=YyVVicZ32M.
64. Xuandong Zhao, Chenwen Liao, Yu-Xiang Wang, and Lei Li. Efficiently identifying watermarked segments in mixed-source texts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6304–6316, Vienna, Austria, July 2025c. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.316. URL https://aclanthology.org/2025.acl-long.316/. ACL Anthology.
65. Chaoyi Zhu, Jeroen Galjaard, Pin-Yu Chen, and Lydia Y Chen. Duwak: Dual watermarks in large language models. In *Findings of the Association for Computational Linguistics*. Association for Computational Linguistics, 2024.
66. George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books, 2016.