CATCH-22*: PARETO FRONTIER FOR DETECTABILITY AND ROBUSTNESS IN LLM WATERMARKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) generate text through probabilistic token sampling, a mechanism increasingly leveraged for inference-time watermarking to verify AI-generated content. As watermarking schemes proliferate, assessing their robustness-detectability trade-off becomes essential to determine whether watermarks can survive output editing while remaining invisible to adversaries. Current evaluation relies on empirical tests lacking provable guarantees. In this work, we present the first information-theoretic framework that rigorously characterizes this fundamental trade-off. We first prove that detectability is determined solely by the sampling strategy, not the model architecture, thereby establishing a hierarchy ranging from undetectable (distribution-preserving) to highly detectable (biased sampling) schemes. Second, we demonstrate an inverse relationship: watermarks robust to text modifications are inherently more detectable by adversaries, creating an irreducible trilemma: no scheme simultaneously achieves high robustness, low detectability, and reliable verification. Motivated by these theoretical constraints, we propose a hybrid watermarking system that adaptively switches sampling strategies based on LLM output edit levels, achieving Pareto-optimal trade-offs. We show that distribution-preserving schemes provide perfect undetectability; however, they are only robust to near-zero adversarial edits. On the other hand, bias-free and biased sampling offer high robustness guarantees at 15-20% output editing, but with detectable output statistics. At high output editing rates, no watermarking provides robustness guarantees. Lastly, we empirically validate our theoretical trade-off claims with Llama 2 7B and Mistral 7B models under paraphrasing attacks, thereby confirming that Pareto-optimality is only achieved by a hybrid watermarking scheme. Overall, our framework provides watermark evaluation beyond empirical testing via principled design, revealing that sampling-based watermarking faces fundamental constraints rooted in information theory rather than implementation limitations.

1 INTRODUCTION

Large Language Models (LLMs) have fundamentally transformed natural language generation, producing text increasingly indistinguishable from human authorship Radford et al. (2019). As these models become ubiquitous in text generation Chung et al. (2024) and summarization Liu & Lapata (2019), they enable malicious applications, including the dissemination of misinformation at scale, contamination of training datasets, and erosion of trust in legitimate AI-generated content. The challenge of distinguishing AI-generated from human-written text has thus become critical Stokel-Walker (2022), with inference-time watermarking emerging as the dominant approach for attribution. However, current watermarking schemes face a fundamental trade-off: robust watermarks that survive text editing introduce detectable statistical artifacts (Gloaguen et al. (2025); Liu et al. (2025)), while provably undetectable watermarks Christ et al. (2024) fail catastrophically under LLM editing as token entropy used to embed the watermark drops Moitra & Golowich (2024).

The rapidly growing class of inference-time LLM watermarking schemes (Fig. 1) employs cryptographic primitives at different stages of token generation to embed verifiable signals in LLM out-

*The name alludes to Joseph Heller's *Catch-22*, a paradoxical dilemma in which one decision cannot be made without negating another. In the context of LLMs, watermarks face an analogous bind: improving robustness often makes them more detectable, while reducing detectability weakens their robustness.

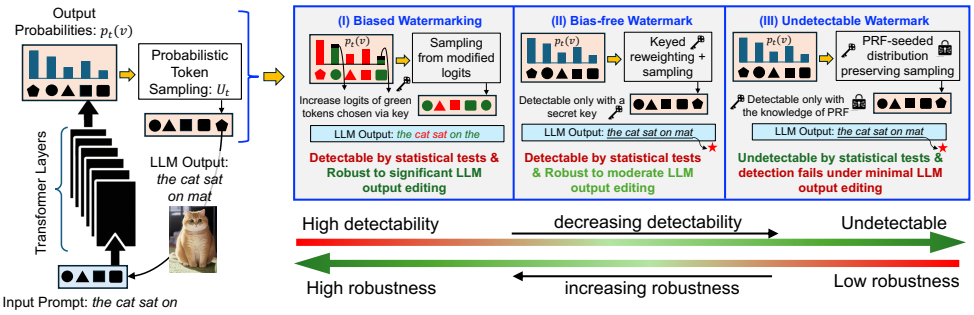


Figure 1: Watermarking schemes in modern LLMs exhibit a trade-off between detectability via statistical tests and robustness against LLM output editing.

puts. Biased sampling methods (Kirchenbauer et al. (2023); Zhao et al. (2023)) use hash functions to designate "green" tokens whose logits are systematically increased, creating detectable statistical signals. Bias-free approaches (Hu et al. (2024); Wu et al. (2024)) employ key-dependent reweighting that preserves expected token distributions while encoding information in variance patterns. Provably undetectable schemes (Christ et al. (2024)) replace sampling randomness with pseudorandom functions (PRFs), achieving perfect undetectability by maintaining exact output distributions. While probability-modifying schemes (biased and bias-free) create redundant statistical signals enabling detection after substantial editing, these deviations are increasingly exposed by black-box statistical tests (Gloaguen et al. (2025)) and targeted prompt analysis (Liu et al. (2025)). Conversely, provable distribution-preserving schemes achieve perfect undetectability but rely on PRF sequences that break under output perturbation, leading to poor robustness Moitra & Golowich (2024).

Although recent work claims provable robustness for undetectable watermarks under bounded edit distance (Moitra & Golowich (2024)), these theoretical guarantees require exponentially large vocabulary sizes that preclude practical deployment. This dichotomy raises a fundamental question: *What is the inherent trade-off between watermark robustness and detectability?*

In this work, we provide a definitive answer through a unified theoretical framework that establishes the fundamental impossibility of simultaneously achieving high robustness, low detectability, and reliable verification. Our analysis reveals that the empirically observed trade-offs (Kirchenbauer et al. (2024); Zhao et al. (2023)) reflect deep information-theoretic constraints rather than limitations of current techniques. Our framework proceeds in two steps: (i) we quantify detectability via total variation distance between watermarked and unwatermarked distributions, proving it depends solely on sampling transformations (Theorem 1), and then (ii) we characterize the information capacity of watermarked LLM outputs under different-editing levels perceived as noise, revealing how capacity determines robustness guarantees (Theorem 2). This framework allows us to ask the question: *What is an optimal watermarking scheme?*

We answer this through the construction of a hybrid watermarking scheme, which selects between probability-modifying and distribution-preserving methods based on noise levels. This hybrid scheme optimizes the watermark parameters to achieve a Pareto-optimal detectability-robustness trade-off (Theorem 3). Experiments with paraphrasing attacks on watermarked outputs from Llama and Mistral models confirm our hybrid scheme achieves a superior trade-off across all noise regimes.

To summarize, our principal contributions are as follows:

1. **Universal detectability bounds:** We establish design-time information-theoretic limits on watermark detectability independent of specific statistical tests or targeted prompt attacks. Detectability remains constant for Greedy sampling, whereas it increases by $O(|\delta|\sqrt{T})$ for biased sampling with bias δ and length T, $O(\sqrt{T})$ for bias-free sampling, while dropping to zero for distribution-preserving schemes (Theorem 1).
2. **Detectability-robustness characterization using information capacity:** We prove that information capacity is inversely related to the detectability. The channel capacity together with the watermark encoding scheme determines robustness guarantees (Theorem 2).
3. **Optimal hybrid watermark construction:** We propose a hybrid watermarking scheme that switches between probability-modifying and distribution-preserving methods based on the noise levels, achieving Pareto-optimal detectability-robustness trade-offs (Theorem 3).

4. **Experimental validation:** We demonstrate the validity of our theoretical predictions through paraphrasing attacks across open-source Llama and Mistral models, confirming that our hybrid scheme achieves Pareto-optimal robustness guarantees even with a 15-20% editing rate, while simultaneously maintaining a total variation distance of < 0.1 compared to unwatermarked outputs.

The remainder of this paper is organized as: Section 2 reviews existing watermarking approaches and their limitations. Section 3 develops our information-theoretic framework, followed by Section 4, which derives the optimal hybrid watermark construction. Section 5 validates our theoretical predictions through comprehensive experiments. Finally, Section 6 concludes the paper.

2 Related Work on LLM Watermarking and Research Gap

Inference-time watermarking for LLMs has evolved rapidly, with schemes progressively trading robustness for undetectability. We categorize existing approaches by their sampling strategies and identify critical gaps that motivate our theoretical framework. Due to space limitations, a comprehensive technical analysis of existing watermarking schemes, along with their corresponding detection schemes, is provided in Appendix A.

Watermarking via Sampling Modifications. Existing watermarking schemes modify the token generation process through three distinct approaches:

1. **Biased sampling** (Kirchenbauer et al. (2023); Zhao et al. (2023)) designates certain tokens as "green" at each step and applies an exponential tilt to the sampling probability. While achieving strong empirical robustness (Kirchenbauer et al. (2024)), these schemes are easily detected through statistical tests (Sadasivan et al. (2023); Gloaguen et al. (2025); Liu et al. (2025)).
2. **Bias-free sampling** (Hu et al. (2024); Wu et al. (2024); Kuditipudi et al. (2024)) employs reweighting functions R_E that preserve expected distributions: $\mathbb{E}_E[R_E(p_t)] = p_t$. Despite maintaining first-order unbiasedness, recent work (Gloaguen et al. (2025)) proves all such schemes remain detectable through variance analysis.
3. **Distribution-preserving sampling**[1] (Christ et al. (2024); Zamir (2024)) provably maintains exact token probabilities ($q_t \equiv p_t$) while replacing true randomness with PRFs: $U_t = \text{PRF}(k, \text{context}_t)$. Though achieving provable undetectability, these schemes fail catastrophically under perturbation to LLM outputs. It is worth noting that, although Moitra & Golowich (2024) proposed a provably undetectable and robust watermarking scheme, it necessitates impractical assumptions, specifically an exponential LLM vocabulary size, to substantiate its robust claims.

This landscape reveals a critical gap: **no existing framework quantifies the fundamental limits of the robustness-detectability trade-off**. Prior works lack: (i) information-theoretic bounds on achievable detectability for given robustness requirements, (ii) analysis revealing why undetectable schemes fail under noise, and (iii) principled construction of schemes that optimally navigate this trade-off. Our work addresses these gaps via an information-theoretic framework, as described in the subsequent sections.

3 Information-theoretic Framework for Robustness vs. Detectability Trade-off Analysis

The detectability and robustness of watermarked text fundamentally depend on how tokens are sampled during generation. When a language model generates text, it proceeds token by token, computing probability distributions over its vocabulary at each step. The actual text produced depends not just on these probabilities but on the sampling rule that converts probabilities into token choices.

[1]Note that we term use *distribution-preserving* for provably undetectable watermarks such as in Christ et al. (2024) unlike statistically indistinguishable watermarks using the same term (Wu et al. (2024)).

Randomness enters this process at each generation step t [2]. The model provides a conditional distribution $p_t(\cdot) = p_\theta(\cdot \mid x, \mathbf{y}_{<t})$ over its vocabulary Σ, where x denotes the initial prompt and $\mathbf{y}_{<t} = (y_1, \ldots, y_{t-1})$ represents the sequence of tokens already generated. To select a token, we need a source of randomness, typically a uniform random variable $U_t \sim \text{Uniform}[0, 1]$. A sampling rule s is a function that takes both p_t and this random variable U_t (possibly along with secret keys) to produce the next token y_t. A watermarked sampling rule modifies either the probabilities (creating $q_t \neq p_t$) or the random variable itself (using a keyed pseudorandom function (PRF)), or both.

Definition 1 (Detectability)**.** *The* detectability *of a watermarking scheme is the statistical distinguishability between watermarked and unwatermarked text. Given a baseline sampling rule s producing distribution P^s over complete texts and a watermarked sampling rule s producing distribution Q^s, the detectability is quantified by the total variation distance:*

$$\text{Detectability}(s) = \text{TV}(P^s, Q^s) = \sup_{A \subseteq \Omega} |P^s(A) - Q^s(A)| \leq \sqrt{\tfrac{1}{2}\,\text{KL}(Q^s \| P^s)}, \tag{1}$$

This measures the maximum distinguishing advantage of any adversary without knowledge of the watermarking key, where the inequality follows from Pinsker's inequality (Pinsker (1964)).

We analyze four sampling approaches spanning the complete spectrum of detectability, building on the watermarking schemes described in Section 2. In addition to the three watermarking approaches, we include **greedy sampling** as a baseline, which eliminates all randomness by always selecting the most probable token: $v_t^\star = \arg\max_v p_t(v)$. Together, these four approaches enable us to characterize how detectability depends on the degree of randomness modification, from complete elimination (greedy) to biased probability adjustments (biased and bias-free sampling) to exact distribution preservation with controlled randomness (distribution-preserving sampling).

3.1 Detectability Characterization

Theorem 1 (Information-theoretic Detectability (single-shot))**.** *Fix a prompt x and length T. Let P^s be the baseline distribution induced by standard stochastic sampling from the model, and let Q denote the distribution induced by a given sampling rule. The* single-shot *total variation (TV) distance between P^s and Q satisfies:*

Sampling Method	***Total Variation Distance***	***Scaling in T***
Greedy	$\text{TV}(P^s, Q^{\text{greedy}}) = 1 - P^s(\mathbf{y}^\star)$	$O(1)$
Biased (δ-tilt)	$\text{TV}(P^s, Q^{\text{bias}_\delta}) \leq \lvert\delta\rvert \sqrt{\frac{1}{4}\sum_{t=1}^{T} g_t(1-g_t)}$	$O(\lvert\delta\rvert\sqrt{T})$
Bias-free (fixed key/code E)	$\text{TV}(P^s, Q_E^{\text{bf}}) \leq \sqrt{\frac{1}{4}\sum_{t=1}^{T}\sum_{v} \frac{\text{Var}_E[R_E(p_t)(v)]}{p_t(v)}}$	$O(\sqrt{T})$
Distribution-preserving (per draw)	$\text{TV}(P^s, Q^{\text{prf}}) = 0$	0

We denote the distributions as follows: (a) Q^{greedy}, which places unit mass on the deterministic greedy sequence $\mathbf{y}^\star$; (b) Q^{bias_δ}, the tilted distribution with bias δ over a keyed green set G_t, where $g_t = p_t(G_t) = \sum_{v \in G_t} p_t(v)$ is its baseline probability mass; (c) Q_E^{bf}, obtained from an unbiased reweighting operator R_E with $\mathbb{E}_E[R_E(p_t)] = p_t$, noting that for a fixed key E one still has $R_E(p_t) \neq p_t$; and (d) Q^{prf}, which preserves $q_t \equiv p_t$ while replacing randomness with PRF coins. The proof is given in Appendix C.

Interpretation of Theorem 1

- **Universal characterization:** Detectability is determined by the sampling transformation alone, independent of model architecture, allowing general analysis via total variation distance.
- **Detectability hierarchy:** Greedy sampling gives constant detectability; biased tilts grow as $O(|\delta|\sqrt{T})$; bias-free reweightings scale as $O(\sqrt{T})$ for a fixed key; and distribution-preserving schemes yield zero single-shot detectability.
- **Accumulation over length:** Any nontrivial deviation from the baseline distribution, however small, accumulates over longer texts.

[2] All the math notations used in this work are described in Appendix B.

The analysis highlights a clear hierarchy: detectability diminishes as sampling rules employ more sophisticated mechanisms, with PRF-based distribution-preserving schemes achieving provable single-shot undetectability (statistical in the random-oracle model and computational with PRFs). This result underpins the robustness–detectability trade-off explored next.

3.2 Robustness analysis under text perturbations

The fundamental tension in watermarking lies in balancing *stealth*, i.e., keeping the generated distribution statistically close to the baseline so unauthorized parties cannot reliably distinguish it [3], with *robustness*, i.e., enabling an authorized key holder to detect the watermark after edits or paraphrasing. We quantify stealth via per-sample KL-divergence drift, and robustness via the detection power of a Neyman–Pearson (NP) test at miss probability β (power $1-\beta$).

Definition 2 (Robustness). *Fix a false-alarm level $\alpha \in (0,1)$. A scheme achieves $(\varepsilon, \alpha, \beta)$-*robustness *on a length-T output if, whenever the edited text $\tilde{y}$ differs from the watermarked text y in at most εT token positions (measured by edit distance* ED*), the key-holder's level-α detector recovers the watermark with miss probability at most β:*

$$\Pr\big[\text{Detect}(k,\tilde{y}) = 1 \,\big|\, \text{ED}(y,\tilde{y}) \le \varepsilon T\big] \;\ge\; 1-\beta. \tag{2}$$

We construct a noise model, where the token at each position is either replaced by a uniformly random element of the LLM vocabulary Σ with probability ε, or left unchanged. For the biased and the bias-free watermarks, we define the induced per-token information at zero edits, D_0 as:

- *Biased*: the sampler tilts toward a key-dependent subset $G \subseteq \Sigma$ of baseline mass $\gamma = \sum_{v\in G} p_t(v)$, producing $q_{t,\delta}(v) \propto p_t(v)\, e^{\delta \mathbf{1}[v\in G]}$ with $D_0^{(\text{biased})} \approx \frac{\delta^2 \gamma(1-\gamma)}{2\ln 2}$.
- *Bias-free*: the sampler reweights by $R_E(v)$ with $\mathbb{E}_E[R_E(v)] = 1$, so $q_{t,E}(v) = p_t(v)R_E(v)$. Therefore we can express $\sigma^2(v) = \text{Var}_E[R_E(v)]$ and $\hat{\sigma}^2 = \sum_v p_t(v)\sigma^2(v)$ leading to $D_0^{(\text{bias-free})} \approx \frac{\hat{\sigma}^2}{2\ln 2}$.

These D_0 values are the noise-free *per-token information budget* available to the optimal NP test. The *total information budget* across T tokens is $\mathsf{TI}(T) := T \cdot D_0$, the natural analogue of blocklength times per-use information in digital communication. Under edits at rate ε, the difference between the watermarked and baseline token distributions at each position is linearly attenuated by $1-\varepsilon$. This is because KL distance is locally quadratic in perturbations, leading to the effective per-token information contracting by $(1-\varepsilon)^2$, so $D_\varepsilon \approx (1-\varepsilon)^2 D_0$. The channel capacity is therefore:

$$C(\varepsilon) := \sum_{t=1}^{T} D(q_{t,\varepsilon} \| p_{t,\varepsilon}) \approx T(1-\varepsilon)^2 D_0. \tag{3}$$

Theorem 2 (Watermark Robustness–Detectability). *We fix T and the substitution channel described above. In the small-signal regime the noise-free per-token information is $D_0^{(\text{biased})} \approx \delta^2\gamma(1-\gamma)/(2\ln 2)$ and $D_0^{(\text{bias-free})} \approx \hat{\sigma}^2/(2\ln 2)$. Under edits at rate ε, the detector's usable information is $C(\varepsilon) \approx T(1-\varepsilon)^2 D_0$. A sufficient condition for power $1-\beta$ in the NP-test is $T(1-\varepsilon)^2 D_0 \ge \log_2(1/\beta)$, making the maximal tolerable edit rate: $\varepsilon_\beta(T, D_0) = 1 - \sqrt{\log_2(1/\beta)/TD_0}$.*

The proof (per-token KL expansions, $(1-\varepsilon)^2$ contraction, chain rule in T) appears in Appendix D.

Interpretation of Theorem 2

- **No single "critical noise" point.** There is no universal edit level where all methods fail. Each watermark has a turning point (knee) determined by the number of tokens examined and the amount of watermark signal placed per token. More tokens or a stronger signal increase the knee value.
- **Total information budget.** The watermark provides a fixed information budget spread across the output. After editing, only a fraction remains, and beyond the knee, no detector can compensate once the budget falls below what is needed for verification.
- **Stealth versus robustness.** If the watermark must stay hard to spot, especially to outsiders who can collect many tokens, the per-token signal must be small. Stronger stealth, therefore, lowers the knee and reduces tolerable editing.

[3] We use *stealth* to mean low *detectability* (i.e., small total variation between the baseline P^s and Q under a fixed prompt and length T) for untrusted parties who do not possess the knowledge of secret key.

3.3 IMPLICATIONS FOR WATERMARK DESIGN

Theorem 2 gives a simple rule of thumb for design: robustness improves by increasing redundancy via number of tokens T and the per-token information budget D_0, and it degrades quadratically with the edit rate through the factor $(1-\varepsilon)^2$. For a given power target β (in NP-test), the operating boundary is a *knee* in ε determined solely by (T, D_0, β).

Watermarking methods that allocate the same total information TD_0 will therefore share the same boundary (w.r.t. to the NP-test), even if they realize the information in different statistical features across LLM outputs. Distribution-preserving (undetectable) schemes sit at the opposite end of this spectrum: because their verification relies on stringent entropy conditions rather than accumulated statistical drift, they are brittle to edits and offer only vanishing tolerance under adversarial perturbations (shown in Appendix D). The following corollary consolidates the baseline impossibility region implied by Theorem 2 and its strengthening when an explicit stealth constraint is imposed.

Corollary 1 (Impossibility Result)**.** *Fix length T, watermark strength D_0 (bits/ token), and target power $1-\beta$. Define the knee*

$$\varepsilon_\beta(T, D_0) \;=\; 1 - \sqrt{\frac{\log_2(1/\beta)}{T \cdot D_0}}. \tag{4}$$

For any $\varepsilon > \varepsilon_\beta(T, D_0)$ one has $T(1-\varepsilon)^2 D_0 < \log_2 \frac{1}{\beta}$, so, beyond this boundary, reliable detection at the specified power is unattainable for any probability-modifying watermark with the given (T, D_0). In particular, seeking high robustness (e.g., $\varepsilon \gtrsim 0.3$) together with strong stealth (small τ for nontrivial M) is incompatible at fixed T.

The proof of the above corollary is provided in Appendix D.9. At a high level, the corollary formalizes the design dilemma: one cannot simultaneously have large edit tolerance, stringent stealth, and guaranteed verification. Practical watermarking must therefore select an operating point along this trade-off, or adopt hybrid schemes that adapt the information budget to the anticipated edit regime while acknowledging the fundamental boundary imposed by $\varepsilon_\beta(T, D_0)$. In the next section, we propose the latter as a Pareto-optimal watermarking scheme in terms of detectability and robustness.

4 CONSTRUCTING OPTIMAL WATERMARKS UNDER OUTPUT EDITING

Building upon the robustness-detectability trade-off in Theorem 2, in this section, we develop a principled construction that finds the optimal watermark parameters based on the edit rate of the output channel. The key idea is that no single family is uniformly optimal across noise regimes. Instead, the operating point should be chosen as a function of the edit rate $\hat{\varepsilon}$, the text length T, and the per-token information budget available to the detector. We refer to the three watermark families: distribution-preserving (DP), bias-free (BF), and biased (B), as described in Section 2, as well as to their detectability behavior (Section 3.1) and small-signal information expansions (Section 3.2).

4.1 A COMPOSITE LOSS FUNCTION

The design objective is to maintain a clear link between stealth and robustness while enabling a clean optimization program. Let $D_0(\theta)$ denote the per-token information (in bits) induced by watermark parameters θ at zero edits. Under the substitution channel, Theorem 2 states that the usable sequence-level signal at edit rate ε equals

$$C(\varepsilon; \theta) \;=\; T(1-\varepsilon)^2\, D_0(\theta), \qquad D_0(\theta) \;\geq\; D_{\mathrm{req}}(\varepsilon, T, \beta) \;:=\; \frac{\log_2(1/\beta)}{T(1-\varepsilon)^2}, \tag{5}$$

which yields a sufficient condition for achieving miss probability at most β with a level-α Neyman–Pearson detector. On the stealth side, Theorem 1 formalizes detectability in terms of total variation for a single shot. We denote the resulting monotone penalty by $\mathrm{TV}_{\mathrm{pen}}(D_0; M)$ for an outsider that can pool M tokens, and we summarize the corresponding stealth cap as $D_{\mathrm{stealth}}(M, \tau)$ for a target TV budget τ. These ingredients motivate an information-aware loss that enforces robustness while discouraging unnecessary statistical drift:

$$\mathcal{L}(\theta; \hat{\varepsilon}, M, \tau) \;=\; \lambda_r \left[\log_2(1/\beta) - T(1-\hat{\varepsilon})^2 D_0(\theta)\right]_+ + \lambda_q\, \mathrm{TV}_{\mathrm{pen}}\big(D_0(\theta); M\big) + \lambda_a\, \mathrm{Amp}(\theta). \tag{6}$$

The hinge in the first term compels the design to supply just enough information to meet the detection requirement in equation 5, and no more. The second term translates the single-shot detectability perspective of Theorem 1 into a conservative, sequence-level penalty that grows monotonically with D_0. The final term regularizes signal amplitude at the parameter level (e.g., $\sqrt{\hat{\sigma}^2}$ for BF and $|\delta|$ for B), thereby favoring parameterizations that realize the same information with smaller perturbations.

4.2 Optimal Watermarking Through Loss Minimization

Minimizing equation 6 reveals a simple and interpretable structure. Because both the detectability penalty and the amplitude penalty increase with D_0, whereas the robustness hinge vanishes once the inequality in equation 5 is met, any minimizer must operate at the smallest feasible per-token information. This observation leads to the target level

$$D^\star \;:=\; \min\Big\{\, D_{\mathrm{stealth}}(M,\tau),\; D_{\mathrm{BF}}^{\max} + D_{\mathrm{B}}^{\max} \,\Big\} \quad \text{subject to} \quad D^\star \;\geq\; D_{\mathrm{req}}(\varepsilon, T, \beta), \tag{7}$$

where $D_{\mathrm{BF}}^{\max}$ and $D_{\mathrm{B}}^{\max}$ denote the small-signal budgets specified previously. If the inequality in equation 7 cannot be satisfied, then the requested power $1-\beta$ is unattainable at the given edit rate under the available stealth and budget constraints.

For a feasible $D^\star$, the remaining decision concerns how to realize this information across the two types of probability-modifying watermarks. Since the TV penalty depends only on $D^\star$ (and not on how it is decomposed), the optimal split minimizes the amplitude term. The family-specific mappings between information and parameters yield a closed-form allocation that prioritizes the bias-free family up to its budget and uses the biased family only for any residual information.

Theorem 3 (Optimal hybrid watermarking)**.** *Fix T, ε, a detector level α, and a power target $1-\beta$. Consider a DP watermark with K marked positions and correction radius t, and the statistical families BF and B with budgets $D_{\mathrm{BF}}^{\max}$ and $D_{\mathrm{B}}^{\max}$ defined earlier. For the loss in equation 6, an optimal strategy $\mathcal{W}^\star(\varepsilon)$ is:*

1. ***DP region (perfect stealth).*** *If the verifier succeeds with probability at least $1-\beta$ under edits (equivalently, if $X \sim \mathrm{Binomial}(K, 1-\varepsilon)$ obeys $\Pr[X < K - t] \leq \beta$ as stated once in the prior section), then DP achieves the target power with $\mathrm{TV} = 0$ and thus minimizes equation 6.*

2. ***Statistical region (information targeting).*** *Otherwise, choose the target information $D^\star$ via equation 7. If $D^\star < D_{\mathrm{req}}(\varepsilon, T, \beta)$, then no watermark can meet the power target at this edit rate.*

3. ***Allocation and parameters.*** *Among all decompositions $D^\star = D_0^{\mathrm{BF}} + D_0^{\mathrm{B}}$ that respect the budgets, the amplitude-minimizing split and corresponding parameter read-off are*

$$D_0^{\mathrm{BF}\star} = \min\{D^\star,\, D_{\mathrm{BF}}^{\max}\}, \qquad \hat{\sigma}^{2\star} = \textit{BF map applied to } D_0^{\mathrm{BF}\star}, \tag{8}$$

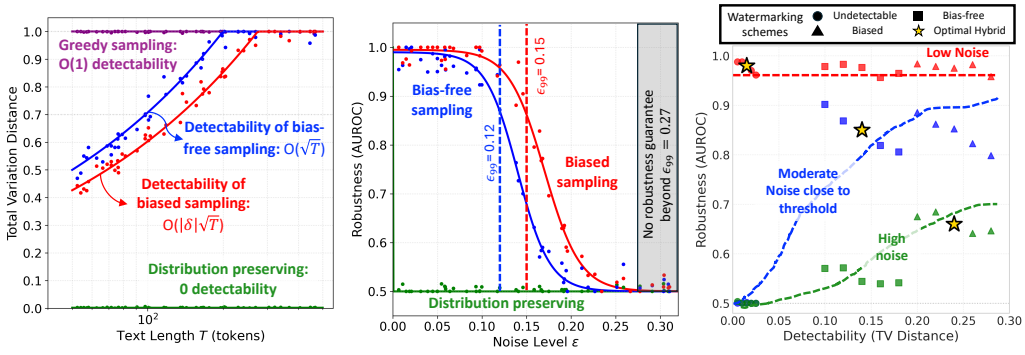
$$D_0^{\mathrm{B}\star} = D^\star - D_0^{\mathrm{BF}\star}, \qquad \delta^\star = \textit{B map applied to } D_0^{\mathrm{B}\star}, \qquad \gamma^\star = \tfrac{1}{2}, \tag{9}$$

where the "BF/B map applied to D_0" refers to $D_0^{(\textit{biased})}$ or $D_0^{(\textit{bias-free})}$ (Section 3.2). In particular, if $D_{\mathrm{req}}(\varepsilon, T, \beta) \leq D_{\mathrm{BF}}^{\max}$, then the optimizer selects a pure BF design; otherwise, BF is saturated and the remainder is realized with B.

The proof of the above theorem is provided in Appendix E.

Interpretation of Theorem 3

1. **No universal optimum.** There is no single watermarking scheme that is best in all situations. When the distribution-preserving verifier succeeds, it should be used because it achieves perfect stealth. Otherwise, a statistical scheme should be selected and tuned to the smallest signal level that still guarantees the target detection power.
2. **Preference for bias-free information.** Among statistical options, the bias-free family is favored first because it achieves the same detection capability with a smaller parameter change. Only when this budget is exhausted should the biased family be used to supply any remaining information.
3. **Limits of feasibility.** If the information required for the desired reliability exceeds what is permitted by stealth constraints and family budgets, then reliable detection cannot be achieved. This identifies a true impossibility region rather than a shortcoming of the detector.



(a) (b) (c)

Figure 2: Empirical validation showing: (a) dependence of total variation (TV) on sampling rule and sequence length, (b) detection AUROC versus edit noise in generated text, and (c) trade-off between attack resistance and detectability across low, moderate, and high noise regimes. The hybrid scheme aligns with the Pareto optimal boundary in every regime.

In summary, the composite loss equation 6 combines the detectability perspective of Theorem 1 with the robustness requirement of Theorem 2 into a single optimization framework. The hinge enforces the minimal information level needed for the desired power, the TV penalty internalizes conservative single-shot detectability into sequence-level design, and the amplitude regularizer privileges parameter-efficient realizations of a fixed information budget. Next, we compare the detectability vs. robustness of our hybrid watermark with other schemes through paraphrasing attacks on LLMs.

5 Experimental Evaluation

This section empirically validates our information-theoretic framework using three families of watermarking schemes, evaluating both detectability and robustness against paraphrasing attacks. All the relevant code for replicating the experiments is available at `https://anonymous.4open.science/r/Catch-22-Pareto-Frontier-Watermark-in-LLMs-040B`. The repository will be made publicly available, ensuring replicability and full functionality, along with detailed user manuals, once the paper is accepted.

Experimental Setup

Dataset and Models. For our non-watermarked baseline, we generate text using 500 prompts randomly sampled from the LFQA dataset, which contains long-form questions from Reddit spanning six domains (July to December 2021) Krishna et al. (2023). We conduct our analysis using open-source Llama-2 7B Touvron et al. (2023) and Mistral 7B Jiang et al. (2023) models on a single NVIDIA H100 GPU, generating outputs ranging from 100 to 1000 tokens.

Watermarking Schemes. We evaluate three categories of watermarking: biased sampling (KGW in Kirchenbauer et al. (2023) and Unigram in Zhao et al. (2023)), bias-free sampling (DiPMark in Wu et al. (2024) and HCW in Hu et al. (2024)), and distribution-preserving sampling (CGZ scheme in Christ et al. (2024)). Additionally, we test our optimal hybrid sampling scheme derived from Theorem 3, which dynamically adapts to observed edit noise levels.

Paraphrasing Attacks. We employ two attack methods: the DIPPER paraphraser in Krishna et al. (2023) with variable token edit rates, and the OPT-2.7B model Zhang et al. (2022) prompted with "Rewrite the following paragraph:", which produces an average edit rate of 15%.

Appendix F, Table 1 provides a comprehensive comparison of robustness versus detectability, demonstrating that our hybrid scheme achieves Pareto optimality across different noise regimes. While this tabular analysis offers a model-agnostic view of the detectability-robustness space, we now present a detailed analysis focusing on Llama 7B outputs subjected to DIPPER paraphrasing at varying edit levels.

5.1 Trade-offs Between Attack Resistance and Detectability

Figure 2(a) demonstrates how total variation (TV) distance scales with output token length, confirming the predictions of Theorem 1. Greedy decoding exhibits $O(1)$ TV scaling with sequence length

T, empirically approaching the upper bound of 1, reflecting the length-independent distributional shift induced by deterministic selection. Biased sampling shows TV growing as $|\delta|\sqrt{T}$, where δ represents bias magnitude. Bias-free sampling displays similar $\sqrt{T}$ growth but with a different constant factor determined by variance modulation rather than mean shifts. Distribution-preserving sampling maintains near-zero TV across all sequence lengths, remaining effectively undetectable. These results validate and formalize previous empirical observations in Kirchenbauer et al. (2024) regarding the improved detectability that comes with increased token length.

Figure 2(b) illustrates detection performance (AUROC) under varying paraphrasing intensities. The curves exhibit a characteristic knee point corresponding to the threshold where the Neyman-Pearson test maintains 99% detection power. The critical noise thresholds $\varepsilon_{99} \approx 0.15$ for biased sampling and 0.12 for bias-free sampling aligns with $T(1-\varepsilon)^2 D_0 \geq \log_2(1/\beta)$ in Theorem 2, with $\beta = 0.01$ and initial information budget $TD_0 = 10$ bits. Below this threshold, both schemes maintain high AUROC, though their degradation patterns differ: bias-free (variance-based) encoding exhibits a sharp decline beyond the knee, while biased (mean-shift) encoding degrades more gradually. Distribution-preserving sampling proves fragile to edits, as its decoding depends on intact high-entropy substrings, rendering it undetectable even under minimal paraphrasing.

Figure 2(c) synthesizes the complete landscape by plotting AUROC against TV distance across three noise regimes, each sampled at five edit rates: low noise (red, $\varepsilon_{99} < 0.005$), moderate noise (blue, $\varepsilon_{99} \approx 0.15$), and high noise (green, $\varepsilon_{99} > 0.15$). No single scheme achieves both undetectability and attack resistance across all regimes. However, the optimal hybrid from Theorem 3 consistently traces the Pareto frontier, crucially outperforming the best existing scheme within each regime. This adaptive approach emerges as the most reliable and stealthy watermarking solution across all noise conditions. By adjusting watermark parameters based on observed edit rates, the hybrid maintains superiority in the AUROC-TV plane, with operating points aligning precisely with theoretical predictions and surpassing any fixed scheme across the entire spectrum of edit intensities. Our framework, therefore, serves as a practical guide for constructing watermarks on the Pareto-optimal frontier of the AUROC-TV plane, as discussed next.

6 Discussion and Conclusion

This work establishes an information-theoretic framework that fundamentally characterizes the trade-off between detectability and edit tolerance in language model watermarks. Biased and bias-free sampling schemes accumulate detectable statistical signals across tokens, enabling reliable recovery under text edits while remaining statistically identifiable. Conversely, distribution-preserving techniques achieve provable undetectability but fail under minimal editing due to their reliance on intact high-entropy patterns. In our analysis, we frame watermark detection as one-bit extraction over a noisy channel, proving that redundancy enhances robustness at the cost of statistical visibility. In other words, this fundamental trade-off cannot be circumvented. Any scheme seeking both properties must compromise on at least one of them. Building on these insights, we develop a hybrid watermarking scheme operating at the Pareto-optimal boundary, consistently outperforming existing approaches across all noise regimes.

This information-theoretic perspective transcends the adversarial cat-and-mouse game of watermarking attacks by providing principled guidance for systematic designers. Rather than pursuing simultaneous robustness and undetectability, designers can adaptively select schemes based on application requirements: deploying undetectable watermarks in privacy-sensitive contexts and robust watermarks in public applications, knowing a single type can't cater to both.

Extensions and Implications. While our analysis focuses on inference-time watermarking, it provides insights for training-time watermarks embedded in model parameters (Appendix F). Since model architecture has a minimal impact on inference-time performance, we believe training-time schemes can potentially exhibit different trade-offs that are worthy of future investigation. Additionally, undetectable watermarking introduces security concerns: the surplus entropy concealing one-bit signals can encode multi-bit payloads, creating covert channels within LLM outputs as proposed in Gaure et al. (2024); Zamir (2024), and we also analyze it further in Appendix G.

All in all, our work identifies Pareto-optimal LLM watermarking solutions and establishes theoretical foundations for practical watermark designs, even when the conflicting goals of high robustness and undetectability cannot be simultaneously achieved.

IMPACT STATEMENT

Practical Deployment. Our work reveals that no watermarking scheme can simultaneously achieve high robustness, strong undetectability, and reliable detection. For controlled environments (enterprise, academic), we recommend undetectable watermarks paired with access controls and key rotation. For public deployments, use detectable watermarks with documented failure modes and regular auditing. System operators should monitor real-world editing patterns and adjust watermarking strategies based on our theoretical thresholds: use distribution-preserving methods for minimal editing ($\varepsilon < 0.05$), variance-based encoding near the critical threshold ($\varepsilon \approx 0.15$), and bias-based methods under heavy editing ($\varepsilon > 0.3$).

Future Work. While our analysis focuses on inference-time watermarking, several directions merit investigation. First, training-time watermarks (Gu et al. (2024)) that embed signals directly into model weights could enable watermarking for open-source models where users control decoding. Key challenges include resistance to fine-tuning attacks and minimizing distillation-induced quality loss. Second, semantic watermarking operating in embedding space may offer orthogonal robustness properties worth characterizing theoretically, such as in images and multimodal data. Finally, the covert channel vulnerability in watermarks (Appendix G) requires further investigation, including the development of detection methods for unauthorized payload embedding.

Limitations. While our framework establishes fundamental bounds for LLM watermarking, it assumes independent token-level editing that sophisticated paraphrasing attacks may violate through correlated changes. However, since the attack on LLMs is an active research area, such paraphrasing attacks are crucial for vulnerability assessment of LLMs, which in turn enhances their security. Additionally, although our hybrid scheme achieves Pareto optimality across noise regimes, it requires accurate estimation of editing levels, which remains challenging in adversarial settings and can be considered as a future direction of research. Nevertheless, our theoretical insights provide essential guidance for practical deployments.

ETHICAL CONSIDERATIONS

We conduct all our experiments on open-source large language models with known vulnerabilities, such as loss of watermarking robustness due to LLM output editing. This research is essential from the perspective of LLM vulnerability assessment, given that these systems are increasingly becoming part of our daily lives. We believe that our theoretical framework and results will assist the research community in designing improved LLM watermarking schemes.

REPRODUCIBILITY

We are firm believers and remain committed to open-source research. The relevant code and its corresponding instructions is available at `https://anonymous.4open.science/r/Catch-22-Pareto-Frontier-Watermark-in-LLMs-040B` for replication of results. This includes models, prompts, watermarking schemes, and paraphrasing attacks to support comparative studies and encourage the community to adopt joint reporting of detectability and robustness of new LLM watermarking schemes.

REFERENCES

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL `https://arxiv.org/abs/2306.09194`.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research (JMLR)*, 25(70):1–53, 2024. URL `https://jmlr.org/papers/volume25/23-0870/23-0870.pdf`.

Simen Gaure, Stefanos Koffas, Stjepan Picek, and Sondre Rønjom. L2m=c large language models are covert channels. *arXiv preprint arXiv:2405.15652*, 2024. URL https://arxiv.org/abs/2405.15652.

Thibaud Gloaguen, Nikola Jovanović, Robin Staab, and Martin Vechev. Black-box detection of language model watermarks. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=E4LAVLXAHW.

Chenchen Gu, Xiang Lisa Li, Percy Liang, and Tatsunori Hashimoto. On the learnability of watermarks for language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=9k0krNzvlV.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=uWVC5FVidc.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023. URL https://arxiv.org/abs/2301.10226.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=DEJIDCmWOz.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In *Proceedings of NeurIPS*, 2023. URL https://arxiv.org/abs/2303.13408.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research (TMLR)*, 2024. URL https://openreview.net/forum?id=FpaCL1MO2C.

Aiwei Liu, Sheng Guan, Yiming Liu, Leyi Pan, Yifei Zhang, Liancheng Fang, Lijie Wen, Philip S Yu, and Xuming Hu. Can watermarked llms be identified by users via crafted prompts? In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=ujpAYpFDEA.

Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*, 2019. URL https://arxiv.org/abs/1908.08345.

Ankur Moitra and Noah Golowich. Edit distance robust watermarks for language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 20645–20693, 2024. URL https://arxiv.org/abs/2406.02633.

Mark S Pinsker. Information and information stability of random variables and processes. *Holden-Day*, 1964.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023. URL https://arxiv.org/abs/2303.11156.

Chris Stokel-Walker. Ai bot chatgpt writes smart essays-should professors worry? *Nature*, 2022. URL https://www.nature.com/articles/d41586-022-04397-7.

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

Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. A resilient and accessible distribution-preserving watermark for large language models. *ICML*, 2024. URL https://openreview.net/pdf?id=c8qWiNiqRY.

Or Zamir. Excuse me, sir? your language model is leaking (information). *arXiv preprint arXiv:2401.10360*, 2024. URL https://arxiv.org/abs/2401.10360.

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. URL https://arxiv.org/pdf/2205.01068.

Xuandong Zhao, Yu-Xiang Wang, and Lei Li. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*, 2023. URL https://arxiv.org/abs/2306.17439.

A Extended Review of LLM Watermarking Literature

We provide here a comprehensive technical analysis of existing watermarking schemes for large language models, extending the overview presented in Section 2. This review organizes prior work according to its fundamental design principles and analyzes its theoretical guarantees, practical limitations, and empirical vulnerabilities.

A.1 Probability-Modifying Watermarks

Probability-modifying watermarks alter token selection probabilities during generation to embed detectable signals. This broad category encompasses all schemes that deviate from the original model's distribution, whether through direct biasing or more subtle statistical modifications.

Biased Sampling Schemes The seminal work of Kirchenbauer et al. (2023) introduced soft watermarking through dynamic vocabulary partitioning. Their scheme computes a cryptographic hash function based on the preceding $k-1$ tokens to partition the vocabulary at each generation step. Specifically, for position t, the vocabulary $\mathcal{V}$ is divided into a green list G_t containing a fraction γ of tokens and a red list $R_t = \mathcal{V} \setminus G_t$. The watermark manifests through logit modification:

$$\hat{\ell}_t[v] = \ell_t[v] + \delta \cdot \mathbf{1}[v \in G_t] \tag{10}$$

where δ controls watermark strength. This induces an exponential tilt in the sampling distribution, increasing the probability of green tokens by approximately a factor e^{δ}. Note that in this work, we use $k-1=1$ preceding tokens when referring to the KGW scheme.

A significant advancement came from Zhao et al. (2023), who demonstrated that fixing the green-red partition across all positions yields superior robustness properties. Their UNIGRAM-WATERMARK scheme establishes tight bounds on output quality degradation through Rényi divergence analysis and proves quantitative robustness guarantees, tolerating $O(n)$ adversarial edits for sequences of length n.

Shortcoming. *Although robust against moderate edits, both KGW and UNIGRAM accumulate outsider evidence at rate $O(|\delta|\sqrt{T})$. Even modest biases create detectable frequency shifts that can be flagged by chi-square tests or amplified by adversarial prompting. Thus, robustness is achieved only at the cost of increased detectability.*

Bias-Free Sampling Schemes While still modifying probabilities, bias-free approaches attempt to preserve expected token distributions through sophisticated reweighting mechanisms. The framework introduced by Hu et al. (2024) employs context-dependent reweighting functions satisfying:

$$\mathbb{E}_E[R_E(p_t)] = p_t \tag{11}$$

where E is a watermark code derived from context and secret key. This ensures the expected distribution over random keys matches the original model's output, though individual samples are drawn from modified distributions.

Similarly, Wu et al. (2024) achieves expectation preservation through vocabulary permutations, while Kuditipudi et al. (2024) employs inverse transform sampling with controlled randomness. All these schemes modify the sampling distribution $q_t \neq p_t$ at each step but maintain $\mathbb{E}[q_t] = p_t$ through careful construction.

Shortcoming. *Despite unbiasedness in expectation, these methods inevitably introduce higher-order variance signatures that grow as $O(\sqrt{T})$. Such distortions are detectable by second-moment tests Gloaguen et al. (2025). To sustain resilience under edits, the injected watermark signal must be amplified, which further undermines stealth. Hence, they cannot simultaneously ensure strong robustness and low detectability.*

A.2 Distribution-Preserving Watermarks

The most recent class of watermarking schemes achieves provable undetectability by maintaining exact output distributions while controlling only the source of randomness.

Cryptographic Undetectability The breakthrough work of Christ et al. (2024) demonstrated that replacing true randomness with pseudorandom functions achieves perfect statistical indistinguishability. Their construction maintains $q_t \equiv p_t$ for all positions while making generation deterministic for key holders. Detection requires exact reproduction of PRF outputs, creating a cryptographic verification mechanism rather than statistical hypothesis testing.

Extensions by Zamir (2024) show that arbitrary payloads can be embedded within this framework by incorporating messages into PRF seeds, enabling covert communication channels with capacity $\Theta(L)$ bits for text length L.

Shortcoming. *While perfectly undetectable in theory ($q_t \equiv p_t$), these schemes collapse under even light paraphrasing. Verification depends on intact PRF alignment, making edit resilience negligible. Attempts to strengthen robustness reintroduce detectable statistical drift, negating their undetectability advantage.*

A.3 Detection Methods and Vulnerabilities

The arms race between watermarking and detection has produced increasingly sophisticated statistical tests that expose subtle artifacts across all scheme categories.

Statistical Detection Methods For probability-modifying watermarks, Sadasivan et al. (2023) demonstrates that simple frequency analysis suffices for detection. Their chi-squared test compares observed versus expected token frequencies:

$$\chi^2 = \sum_{v \in \mathcal{V}} \frac{(f_v^{\text{obs}} - f_v^{\text{exp}})^2}{f_v^{\text{exp}}} \tag{12}$$

where f_v denotes the frequency of token v. This test achieves high power against biased watermarks with modest sample sizes.

For expectation-preserving schemes, Gloaguen et al. (2025) develops second-moment tests that detect variance anomalies. Their test statistic aggregates squared deviations from expected variance:

$$T = \sum_{t=1}^{n} \left(\|\hat{p}_t\|_2^2 - \mathbb{E}[\|p_t\|_2^2] \right) \tag{13}$$

This approach succeeds because reweighting necessarily introduces variance distortions even when preserving expectations.

Adaptive Attacks Beyond passive detection, Liu et al. (2025) demonstrates active attacks using adversarial prompting. By crafting prompts that amplify watermark biases, they force watermarked models to produce highly distinguishable outputs. Their optimization finds prompts maximizing:

$$\Delta(x) = \mathbb{E}_{y \sim \hat{p}(\cdot|x)}[\text{score}(y)] - \mathbb{E}_{y \sim p(\cdot|x)}[\text{score}(y)] \tag{14}$$

where score measures watermark strength. Such targeted attacks reduce required sample sizes by orders of magnitude.

Note. *These detection methods and attacks highlight a structural vulnerability: biased schemes are easily exposed via frequency analysis, bias-free schemes via variance anomalies, and both via adversarial prompting. Thus, neither family achieves low detectability in practice.*

A.4 Alternative Approaches to Watermark Robustness Analysis

While our main analysis models text perturbations as a noisy channel, several alternative mathematical frameworks have been developed in the watermarking literature to analyze robustness. We review two prominent approaches here.

Direct Statistical Analysis of Detection Scores

Zhao et al. (2023) analyzed robustness by directly tracking how the watermark detection statistic degrades under edit operations. Their approach does not invoke channel capacity or information-theoretic arguments, but instead provides explicit bounds on the z-score used for detection.

For their UNIGRAM-WATERMARK scheme, they prove that if text y is watermarked and an adversary produces modified text u with edit distance $\eta = \text{ED}(y, u) < n$, then the detection z-scores satisfy:

$$z_u \geq z_y - \max\left\{\frac{(1+\gamma/2)\eta}{\sqrt{n}}, \frac{(1-\gamma/2)\eta}{\sqrt{n-\eta}}\right\} \tag{15}$$

where γ is the green list ratio parameter, the proof technique involves analyzing how each edit operation affects the count of green tokens, using a Taylor expansion argument to bound the worst-case degradation. This approach yields that the watermark can tolerate up to $O(n)$ arbitrary edits for text of length n when the watermark strength parameter δ is constant.

The key advantage of this direct approach is its simplicity and explicitness; it provides concrete formulas for how detection degrades with edits. However, it is specific to their particular watermarking scheme and does not readily generalize to other watermarking methods.

Shortcoming. *Although offering concrete edit tolerance formulas, this method is tied to UNIGRAM and does not generalize. Moreover, it provides no guarantees about detectability, limiting its applicability for designing low-detectability watermarks.*

Coding-Theoretic Constructions with Indexing

Moitra & Golowich (2024) took a fundamentally different approach by explicitly constructing watermarks using error-correcting codes. Their key innovation is the concept of indexing pseudorandom codes, which enables robustness to insertions and deletions in addition to substitutions. The construction begins with a binary pseudorandom code (PRC) that is robust to substitutions, then creates an "indexing PRC" over a larger alphabet whose size is a multiple of the original code length. Each symbol in this larger alphabet maps to an index position via a random function, with multiple symbols mapping to each index to provide redundancy.

During encoding, they generate a binary codeword and output symbols whose indices correspond to the positions of ones in that codeword. For decoding, received symbols are mapped back to indices to reconstruct a binary string, which is then decoded using the original binary PRC. The redundancy parameter is crucial for handling insertions and deletions—when an adversary makes edits, the set of indices changes, but the redundancy ensures that with high probability, the Hamming distance between the original and modified binary strings remains bounded.

Their analysis proves that this watermarking scheme achieves substring robustness, meaning that any sufficiently high-entropy substring of watermarked text remains detectable even after a constant fraction of edits. The scheme tolerates a fraction of edits that scales quadratically with the entropy rate, though it requires the alphabet size to grow polynomially with the security parameter.

Shortcoming. *The required exponential vocabulary growth (e.g., n^{23} for 10% edits, where n is the block length) makes the method impractical for real-world LLMs, where vocabularies are capped at ∼30k–100k tokens.*

B NOTATION AND VARIABLES

NOTATION CONVENTIONS

- **Vectors:** $\mathcal{V}^T$ denotes T-length token sequences from vocabulary $\mathcal{V}$.
- **Subscripts:** t indexes token position (1 to T).
- **Superscripts:** On Q indicate sampling method; asterisk ($*$) denotes optimal values.
- **Context:** Conditionals like $p_t(\cdot)$ depend on $\mathbf{y}_{<t}$ and prompt x.
- **Overloading note:** M denotes outsider pooled tokens in §4.1; in Appendix G it denotes the number of covert messages.

CORE VARIABLES AND DISTRIBUTIONS

Symbol	Type/Dim	Description	Sections
L, T	Scalar	Text length (number of tokens)	§3, §4
$\mathcal{V}, \Sigma$	Set	Token vocabulary	§3
x	Vector	Initial prompt	§3
$\mathbf{y} = (y_1, \ldots, y_T)$	$\mathcal{V}^T$	Generated token sequence	§3
$\mathbf{y}_{<t}$	$\mathcal{V}^{t-1}$	Tokens before position t	§3
$\tilde{\mathbf{y}}$	$\mathcal{V}^T$	Edited/noisy text	§3
$\mathbf{y}^\star$	$\mathcal{V}^T$	Deterministic greedy path	§3
$p_t(\cdot)$, $p_\theta(\cdot\|\cdot)$	Function	Baseline LLM conditional probabilities	§3
$q_t(\cdot)$	Function	Watermarked conditional probabilities	§3
P^s	Distribution	Baseline sampling distribution over sequences	§3
$Q^{\mathcal{W}}$	Distribution	Sequence distribution for scheme $\mathcal{W}$	§3
Q^{greedy}	Distribution	Greedy sampling distribution	§3
Q^{bias_δ}	Distribution	Biased (tilted) sampling with parameter δ	§3
Q_E^{bf}	Distribution	Bias-free sampling with key/code E	§3
Q^{prf}	Distribution	PRF-based distribution-preserving sampling	§3
U_t	$[0, 1]$	Uniform random variable used for sampling	§3
U	$\Delta(\Sigma)$	Uniform distribution on Σ	App. D
$T_\varepsilon(P)$	Operator	Edit channel: $(1-\varepsilon)P + \varepsilon U$	App. D
$p_{t,\varepsilon}$, $q_{t,\varepsilon}$	Function	Edited conditionals: $T_\varepsilon(p_t)$, $T_\varepsilon(q_t)$	App. D

WATERMARKING PARAMETERS

Symbol	Type/Dim	Description	Sections
$\delta, \delta^\star$	Scalar	Bias strength (optimal value $\delta^\star$)	§3, §4
$G_t \subset \mathcal{V}$	Set	Keyed green token set at step t	§3
$g_t = p_t(G_t)$	$[0, 1]$	Baseline green mass at step t	§3, App. C
$\gamma, \gamma^\star$	$[0, 1]$	Typical/target green mass (often $\gamma^\star = \frac{1}{2}$)	§3, §4
k	Key	Secret cryptographic key	§3
E, E_t	Code	Keyed code or permutation for bias-free schemes	§3
R_E	Function	Reweighting operator with $\mathbb{E}_E[R_E(p_t)] = p_t$	§3
$\sigma^2(v)$, $\hat{\sigma}^2$	Scalar	$\sigma^2(v) = \text{Var}_E[R_E(p_t)(v)]$, $\hat{\sigma}^2 = \sum_v p_t(v)\sigma^2(v)$	§3, App. D
Z_t	Scalar	Normalizer for tilted sampling	App. C
PRF	Function	Pseudorandom function for RNG replacement	§3
$\mathcal{W}, \mathcal{W}^\star(\varepsilon)$	Scheme	Watermarking scheme and the optimal hybrid	§4, App. E
K, t	Scalars	DP verifier: marked positions K and correction radius t	§4, App. E

INFORMATION THEORY AND ROBUSTNESS

Symbol	Type/Dim	Description	Sections
D_0, D_ε	Bits/token	Per-token information at 0 edits and at rate ε	§3, App. D
$C(\varepsilon)$	Bits	Total usable information $\approx T(1-\varepsilon)^2 D_0$	§3, App. D
$\varepsilon_\beta(T, D_0)$	$[0,1]$	"Knee": $1-\sqrt{\log_2(1/\beta)/(TD_0)}$	App. D
$\mathrm{TV}(P,Q)$	$[0,1]$	Total variation distance	§3, App. C
$\mathrm{KL}(Q\|P)$	$[0,\infty)$	Kullback–Leibler divergence (base 2 in proofs)	§3
$H(\cdot)$, $H_2(\cdot)$	Function	Entropy, binary entropy	§3
$I(\cdot;\cdot)$	Bits	Mutual information	App. G
$\mathrm{Detectability}(s)$	$[0,1]$	Distinguishability for sampling rule s	§3
ε, $\hat{\varepsilon}$	$[0,1]$	Edit rate (true and estimated)	§3, §4
α, β	$[0,1]$	Detector level and miss probability (power $= 1-\beta$)	§4, App. D
$D_{\mathrm{req}}(\varepsilon, T, \beta)$	Bits/token	$\log_2(1/\beta)/T(1-\varepsilon)^2$	§4.1, App. D
M, τ	Scalar, $[0,1]$	Outsider pooled tokens M and TV budget τ	§4.1, App. D
$D_{\mathrm{stealth}}(M,\tau)$	Bits/token	Stealth cap $\frac{2\tau^2}{M\ln 2}$	§4.1, App. D
z, $z_{\mathrm{threshold}}$	Scalar	Z-score statistic and threshold	App. D
N_{green}	Scalar	Count of green tokens	App. D
$\Phi(\cdot)$, $\Phi^{-1}(\cdot)$	Function	Standard normal CDF and its inverse	§4

OPTIMIZATION AND OPERATORS

Symbol	Type/Dim	Description	Sections
$\mathcal{L}(\theta; \hat{\varepsilon}, M, \tau)$	Scalar	Composite loss	§4.1
θ	Variable	Scheme parameters	§4
λ_r, λ_q, λ_a	Scalars	Weights for reliability, stealth penalty, amplitude	§4.1
$D^\star$	Bits/token	Target per-token information after constraints	§4.2, App. E
$D_{\mathrm{BF}}^{\max}$, $D_{\mathrm{B}}^{\max}$	Bits/token	Available budgets for BF and B families	§4.2, App. E
$\mathrm{TV}_{\mathrm{pen}}(D_0; M)$	Scalar	Monotone detectability penalty used in the loss	§4.1
$\mathrm{Amp}(\theta)$	Scalar	Amplitude regularizer (e.g., $\sqrt{\hat{\sigma}^2}$ or $\lvert\delta\rvert$)	§4.1
$\mathbb{E}[\cdot]$, $\mathrm{Var}[\cdot]$	Operator	Expectation, variance	§3
$\mathbf{1}[\cdot]$	Function	Indicator	§3
arg max, sup	Operator	Maximizer, supremum	§3
ln, log, $\log_2$	Function	Natural log, log, base-2 log	§3
$O(\cdot)$, $o(\cdot)$, $\Theta(\cdot)$, $\omega(\cdot)$, $\Omega(\cdot)$	Notation	Asymptotic notation	§3
$\approx$	Operator	Approximately equal	App. D
∞	Symbol	Infinity	App. E

ADDITIONAL SYMBOLS USED IN APPENDIX G

Symbol	Type/Dim	Description	Sections
W	RV	Message index (uniform over $\{1,\dots,M\}$)	App. G
Q_w, Q	Distribution	Distribution for message w and outsider mixture $\frac{1}{M}\sum_w Q_w$	App. G
$C_\star$	Scalar	Mixture divergence budget $D(Q\|P) \le C_\star$	App. G
θ (Appendix)	$[0,1]$	Activity probability $c/\sqrt{L}$ in square-root law construction	App. G
c, κ	Scalar	Constants in square-root law achievability	App. G

C PROOF OF THEOREM 1

This appendix proves the bounds in Theorem 1 and provides a detailed explanation of each step. The statement in the main paper is for computing the total variation distance from a *single-shot* or a single generated text from LLM. Multi-shot black-box detection over multiple LLM queries, key-averaged (n-shot) properties for bias-free watermarks, as well as the computational undetectability guarantees for PRF-seeded schemes, are also described later in this proof.

We measure distributional separation with the total variation distance

$$\mathrm{TV}(P,Q) = \frac{1}{2}\sum_{\mathbf{y}} \big|P(\mathbf{y}) - Q(\mathbf{y})\big|, \tag{16}$$

and we use the Kullback–Leibler (KL) divergence

$$\mathrm{KL}(Q\|P) = \mathbb{E}_{\mathbf{y}\sim Q}\left[\log \frac{Q(\mathbf{y})}{P(\mathbf{y})}\right]. \tag{17}$$

Pinsker's inequality connects these two quantities and will be invoked repeatedly:

$$\mathrm{TV}(P,Q) \le \sqrt{\frac{1}{2}\,\mathrm{KL}(Q\|P)}. \tag{18}$$

For autoregressive distributions that factorize across positions, the KL chain rule expresses the sequence level divergence as a sum of conditional one-step divergences:

$$\mathrm{KL}(Q\|P) = \sum_{t=1}^{T} \mathbb{E}_{y_{<t}\sim Q}\Big[\mathrm{KL}\big(q_t(\cdot \mid y_{<t}) \,\|\, p_t(\cdot \mid y_{<t})\big)\Big]. \tag{19}$$

C.1 Greedy sampling

Let Q^{greedy} be the degenerate distribution that puts unit mass on the unique greedy path $\mathbf{y}^\star = (y_1^\star, \ldots, y_T^\star)$, with $y_t^\star = \arg\max_v p_t(v \mid y_{<t}^\star)$. Since $Q^{\mathrm{greedy}}(\mathbf{y}^\star) = 1$ and $Q^{\mathrm{greedy}}(\mathbf{y}) = 0$ for all $\mathbf{y} \neq \mathbf{y}^\star$, the total variation distance expands as

$$\mathrm{TV}(P^s, Q^{\mathrm{greedy}}) = \frac{1}{2}\sum_{\mathbf{y}} \left|P^s(\mathbf{y}) - Q^{\mathrm{greedy}}(\mathbf{y})\right| \tag{20}$$

$$= \frac{1}{2}\left(\left|P^s(\mathbf{y}^\star) - 1\right| + \sum_{\mathbf{y}\neq\mathbf{y}^\star}\left|P^s(\mathbf{y}) - 0\right|\right). \tag{21}$$

The absolute value in the first term simplifies to $1 - P^s(\mathbf{y}^\star)$ because probabilities are at most one. The sum over the remaining sequences simplifies to $\sum_{\mathbf{y}\neq\mathbf{y}^\star} P^s(\mathbf{y}) = 1 - P^s(\mathbf{y}^\star)$ because the probabilities must sum to one. Therefore

$$\mathrm{TV}(P^s, Q^{\mathrm{greedy}}) = 1 - P^s(\mathbf{y}^\star). \tag{22}$$

C.2 Biased sampling (exponential tilt, soft green list)

At position t, let $G_t \subseteq \mathcal{V}$ denote the keyed green set and define its baseline probability mass $g_t := p_t(G_t) = \sum_{v\in G_t} p_t(v)$. The biased sampler applies an exponential tilt to tokens in G_t:

$$q_t(v) = \frac{p_t(v)\exp\{\delta\,\mathbf{1}[v \in G_t]\}}{Z_t}, \qquad Z_t = \sum_v p_t(v)\exp\{\delta\,\mathbf{1}[v \in G_t]\}. \tag{23}$$

The normalizer follows from splitting the sum into green and non-green tokens. The mass of the complement of G_t is $1 - g_t$ and the mass of G_t is g_t, hence

$$Z_t = (1 - g_t) + g_t e^\delta = 1 + g_t\,(e^\delta - 1). \tag{24}$$

The one-step KL divergence equals

$$\mathrm{KL}(q_t\|p_t) = \sum_v q_t(v)\log\frac{q_t(v)}{p_t(v)} \tag{25}$$

$$= \sum_v q_t(v)\,(\delta\,\mathbf{1}[v \in G_t] - \log Z_t) \tag{26}$$

$$= \delta\, q_t(G_t) - \log\big(1 + g_t(e^\delta - 1)\big). \tag{27}$$

The second line uses the explicit tilted form of q_t, which cancels the factor $p_t(v)$ and yields a term that depends only on Z_t. The last line replaces the indicator sum by the mass $q_t(G_t)$.

A small parameter expansion provides an explicit constant. Using $e^\delta = 1 + \delta + \frac{\delta^2}{2} + O(\delta^3)$ and $\log(1+x) = x - \frac{x^2}{2} + O(x^3)$, the logarithm of the normalizer expands as

$$\log Z_t = \log\Big(1 + g_t\left(\delta + \tfrac{\delta^2}{2} + O(\delta^3)\right)\Big) \tag{28}$$

$$= g_t\delta + \frac{g_t(1-g_t)}{2}\,\delta^2 + O(\delta^3). \tag{29}$$

In the previous step, since g_t is a fixed probability mass, it is absorbed in the last term $O(\delta^3)$. The mass of the green set under q_t is a ratio of two series. Using the series for e^δ and the identity $(1+u)^{-1} = 1 - u + O(u^2)$ with $u = g_t(\delta + \frac{\delta^2}{2}) + O(\delta^3)$ gives

$$q_t(G_t) = \frac{g_t e^\delta}{1 + g_t(e^\delta - 1)} = g_t + g_t(1-g_t)\delta + O(\delta^2). \tag{30}$$

Substituting both expansions into the one-step KL cancels the linear terms and leaves the quadratic coefficient

$$\mathrm{KL}(q_t \| p_t) = \frac{g_t(1-g_t)}{2}\,\delta^2 + O(\delta^3). \tag{31}$$

At the sequence level, the chain rule equation 19 expresses the KL divergence as a sum of the conditional one-step terms under the biased process. Keeping the leading order in δ yields

$$\mathrm{KL}\big(Q^{\mathrm{bias}_\delta} \parallel P^s\big) = \frac{\delta^2}{2}\sum_{t=1}^{T} \mathbb{E}_{y_{<t}\sim Q^{\mathrm{bias}_\delta}}\big[g_t(1-g_t)\big] + O(T|\delta|^3). \tag{32}$$

Finally, Pinsker's inequality converts this to a total variation bound,

$$\mathrm{TV}(P^s, Q^{\mathrm{bias}_\delta}) \le |\delta|\sqrt{\frac{1}{4}\sum_{t=1}^{T}\mathbb{E}[g_t(1-g_t)]} + O\big(\sqrt{T}\,|\delta|^{3/2}\big), \tag{33}$$

which exhibits the $O(|\delta|\sqrt{T})$ scaling with an explicit leading constant.

C.3 Bias-free sampling (unbiased reweighting)

In the bias free setting a keyed operator $R_E : \Delta(\mathcal{V}) \to \Delta(\mathcal{V})$ reweights the baseline, and unbiasedness requires $\mathbb{E}_E[R_E(p_t)] = p_t$ for every step. For a fixed key E one can write

$$R_E(p_t)(v) = p_t(v) + \epsilon_t^{(E)}(v), \qquad \sum_v \epsilon_t^{(E)}(v) = 0, \tag{34}$$

where the sum constraint enforces normalization. The one-step KL divergence admits a Taylor expansion around p_t:

$$\mathrm{KL}(R_E(p_t) \parallel p_t) = \sum_v \big(p_t(v) + \epsilon_t^{(E)}(v)\big)\log\left(1 + \frac{\epsilon_t^{(E)}(v)}{p_t(v)}\right) \tag{35}$$

$$= \sum_v \left[\epsilon_t^{(E)}(v) + \frac{1}{2}\frac{\big(\epsilon_t^{(E)}(v)\big)^2}{p_t(v)} + O\left(\frac{|\epsilon_t^{(E)}(v)|^3}{p_t(v)^2}\right)\right]. \tag{36}$$

The second line follows from $\log(1+u) = u - \frac{u^2}{2} + O(u^3)$ with $u = \epsilon_t^{(E)}(v)/p_t(v)$, distributing the factor $p_t(v) + \epsilon_t^{(E)}(v)$ and combining like terms. The linear term sums to zero if one averages over keys because $\mathbb{E}_E[\epsilon_t^{(E)}(v)] = 0$ by unbiasedness. Therefore, taking the expectation over E yields

$$\mathbb{E}_E[\mathrm{KL}(R_E(p_t) \parallel p_t)] = \sum_v \frac{\mathrm{Var}_E[R_E(p_t)(v)]}{2\,p_t(v)} + o(\|\epsilon\|^2). \tag{37}$$

Summing across positions with the chain rule equation 19 and applying Pinsker's inequality leads to the single-shot bound for a fixed key

$$\mathrm{KL}\big(Q_E^{\mathrm{bf}} \parallel P^s\big) = \sum_{t=1}^{T} \mathrm{KL}(R_E(p_t) \parallel p_t)\,, \tag{38}$$

$$\mathrm{TV}\big(P^s, Q_E^{\mathrm{bf}}\big) \le \sqrt{\frac{1}{4}\sum_{t=1}^{T}\sum_{v}\frac{\mathrm{Var}_E[R_E(p_t)(v)]}{p_t(v)}}, \tag{39}$$

which shows the $O(\sqrt{T})$ scaling and makes explicit the variance controlled constant. This is the detector's view with a fixed key. For completeness, we record a separate mixture view. If the implementation guarantees fresh, independent codes across positions and queries by maintaining a context code history that forbids reuse, then the joint distribution averaged over keys coincides with the baseline for any finite number of generations, which is often referred to as n shot undetectability. That statement concerns a mixture of keys and is distinct from the fixed key detectability bound developed above.

C.4 Distribution preserving sampling (per draw)

If a keyed pseudorandom source replaces randomness while the per-step probabilities remain unchanged, that is $q_t \equiv p_t$ for all histories, then the induced sequence distribution equals the baseline:

$$Q^{\mathrm{prf}}(y_{1:T}) = \prod_{t=1}^{T} q_t(y_t \mid y_{<t}) = \prod_{t=1}^{T} p_t(y_t \mid y_{<t}) = P^s(y_{1:T}). \tag{40}$$

Consequently

$$\mathrm{TV}(P^s, Q^{\mathrm{prf}}) = \frac{1}{2}\sum_{\mathbf{y}} \big|P^s(\mathbf{y}) - Q^{\mathrm{prf}}(\mathbf{y})\big| = 0. \tag{41}$$

This identity formalizes the intuitive fact that sampling from the same conditional laws produces the same sequence distribution, independent of how the coins are generated, as long as they are fresh and independent at each step. □

D Proof of Theorem 2

This appendix provides a complete derivation of Theorem 2. Throughout the appendix, *all logarithms are base 2*, so KL divergences and mutual informations are measured in *bits*. We write $D(P\|Q)$ for the Kullback–Leibler (KL) divergence between distributions P and Q.

We model watermark verification as a binary hypothesis test. The null hypothesis H_0 corresponds to unwatermarked text generated by the baseline sampler, whereas the alternative H_1 corresponds to text produced by a watermarked sampler. Formally,

$$H_0:\ \text{unwatermarked text} \qquad \text{vs.} \qquad H_1:\ \text{watermarked text}, \tag{42}$$

where the observation is taken *after* a perturbation channel that edits tokens independently with rate $\varepsilon \in [0, 1]$.

We adopt a single perturbation model used consistently throughout. Let Σ denote the vocabulary. At each token position $t \in \{1, \dots, L\}$, the edited token $\tilde{Y}_t$ is drawn as

$$\tilde{Y}_t \;=\; \begin{cases} Y_t, & \text{with probability } 1-\varepsilon, \\ U_t, & \text{with probability } \varepsilon, \end{cases} \qquad U_t \sim \mathrm{Uniform}(\Sigma) \text{ and independent of everything else.} \tag{43}$$

Equivalently, if P is a distribution on Σ, the edit channel acts as a convex combination

$$T_\varepsilon(P) \;=\; (1-\varepsilon)\,P \;+\; \varepsilon\,U, \quad \text{with } U \text{ uniform on } \Sigma. \tag{44}$$

Thus the pre-noise per-position conditionals $p_t(\cdot)$ (baseline) and $q_t(\cdot)$ (watermarked) are mapped to $p_{t,\varepsilon} = T_\varepsilon(p_t)$ and $q_{t,\varepsilon} = T_\varepsilon(q_t)$.

The detection problem is posed at a fixed false-alarm level α. We write β for the miss probability (so power is $1-\beta$). The central idea is that a sufficient condition for achieving a given β is that the *total* KL divergence from H_1 to H_0 on the observed sequence exceeds $\log_2(1/\beta)$. This is captured by a Stein-type sufficient condition (Lemma 2). To use it, we (i) quantify the per-token information contributed by the watermark at zero edits, (ii) show how this information contracts under the edit channel, and (iii) aggregate across the sequence by the KL chain rule.

We analyze two small-signal watermark families. In the *biased* (green-list) family, the watermarker tilts the baseline distribution towards a key-dependent subset $G \subseteq \Sigma$ with baseline mass $\gamma = \sum_{v\in G} p_t(v)$. Writing the tilt parameter as δ,

$$q_{t,\delta}(v) \;\propto\; p_t(v)\, e^{\delta\, \mathbf{1}[v\in G]}. \tag{45}$$

A Taylor expansion shows that the corresponding per-token KL is quadratic in δ. In the *bias-free* (variance) family, the watermarker reweights by $R_E(v)$ with $\mathbb{E}_E[R_E(v)] = 1$, i.e.,

$$q_{t,E}(v) \;=\; p_t(v)\, R_E(v), \tag{46}$$

and the per-token KL is quadratic in the reweighting variance. The following lemmas formalize these statements and prepare the ground for the edit-channel analysis.

D.1 Preliminaries: KL expansions and a reliability bound

The first lemma is a standard second-order expansion of KL divergence around a reference distribution, with an explicit remainder bound. It formalizes that, locally, KL equals a quadratic form (the Fisher information metric) up to third-order terms.

Lemma 1 (Second-order KL expansion around p)**.** *Let p be a distribution on a finite set and $q = p+r$ for some perturbation r with $\sum_v r(v) = 0$ and $\|r\|_\infty \le \eta < \min_v p(v)$. Then*

$$D(q\|p) \;=\; \frac{1}{2\ln 2}\sum_v \frac{r(v)^2}{p(v)} \;+\; R, \quad \textit{with } |R| \;\le\; \frac{C}{\ln 2}\,\|r\|_\infty \sum_v \frac{r(v)^2}{p(v)} \tag{47}$$

for an absolute constant C. In particular, when $\|r\|_\infty \to 0$,

$$D(q\|p) = (1+o(1))\;\frac{1}{2\ln 2}\sum_v \frac{r(v)^2}{p(v)}. \tag{48}$$

Proof. Write

$$D(q\|p) = \sum_v (p(v)+r(v))\log\!\left(1+\frac{r(v)}{p(v)}\right). \tag{49}$$

Set $x_v := r(v)/p(v)$. By Taylor's theorem with remainder for $\log(1+x)$,

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3(1+\theta x)^3} \tag{50}$$

for some $\theta = \theta(x) \in (0,1)$ when $|x| < 1$. Using this with $x = x_v$ and noting $\sum_v r(v) = 0$,

$$D(q\|p) = \sum_v (p(v)+r(v))\left(x_v - \frac{x_v^2}{2} + \frac{x_v^3}{3(1+\theta_v x_v)^3}\right) \tag{51}$$

$$= \sum_v \left(p(v)x_v - \frac{p(v)x_v^2}{2}\right) \;+\; \sum_v r(v)x_v \;+\; \sum_v (p(v)+r(v))\frac{x_v^3}{3(1+\theta_v x_v)^3}. \tag{52}$$

The first sum simplifies to $-\frac{1}{2}\sum_v r(v)^2/p(v)$. The second sum equals $\sum_v r(v)^2/p(v)$. Combining these two gives

$$\frac{1}{2}\sum_v \frac{r(v)^2}{p(v)}. \tag{53}$$

For the remainder, since $|x_v| \le \|r\|_\infty/p_{\min} =: \tau < 1$, we have $\left|(1+\theta_v x_v)^{-3}\right| \le (1-\tau)^{-3}$ and $|p(v)+r(v)| \le p(v) + \|r\|_\infty$. Therefore

$$\left|\sum_v (p(v)+r(v))\frac{x_v^3}{3(1+\theta_v x_v)^3}\right| \;\le\; \frac{1}{3(1-\tau)^3}\sum_v \big(p(v)+\|r\|_\infty\big)\frac{|r(v)|^3}{p(v)^3}. \tag{54}$$

Applying the crude bound $|r(v)| \le \|r\|_\infty$ and $p(v) \ge p_{\min}$ yields

$$|R| \;\le\; \frac{C'}{\ln 2}\,\|r\|_\infty \sum_v \frac{r(v)^2}{p(v)} \tag{55}$$

for a constant C' depending only on $p_{\min}$ and τ; absorbing constants gives the stated bound with C. Dividing by $\ln 2$ converts from nats to bits. The $o(1)$ claim follows as $\|r\|_\infty \to 0$. □

The second lemma provides the reliability criterion we will use to translate available information into detection power. It asserts that, for independent per-token contributions, having total KL at least $\log_2(1/\beta)$ is sufficient to drive the miss probability below β at fixed false-alarm level α. We present a standard achievability proof based on the Neyman–Pearson (NP) test, an exponential Markov bound to control the level, and a Cramér–Chernoff bound (in base 2) under the alternative to control the miss probability.

Lemma 2 (Stein's sufficient condition (bits form))**.** *Consider a simple binary hypothesis test between product distributions on sequences of length L, or more generally conditionals whose log-likelihood ratio is a sum of independent terms with finite moment generating function in a neighborhood of the origin. For any level $\alpha \in (0,1)$ and any $\beta \in (0,1)$, there exists $L_0(\alpha,\beta)$ such that for all $L \ge L_0$ the NP test with threshold chosen to achieve level at most α has miss probability at most β whenever*

$$\sum_{t=1}^{L} D\big(P_t^{(1)} \| P_t^{(0)}\big) \;\ge\; \log_2 \frac{1}{\beta} \;+\; o(L). \tag{56}$$

In particular, ignoring the lower-order $o(L)$ term yields the clean sufficient rule $\sum_{t=1}^{L} D(P_t^{(1)} \| P_t^{(0)}) \ge \log_2(1/\beta)$.

Proof. Let $Z_t = \log_2\big(\frac{P_t^{(1)}(Y_t)}{P_t^{(0)}(Y_t)}\big)$ and $S_L = \sum_{t=1}^{L} Z_t$ be the base-2 log-likelihood ratio (LLR) of the sequence. The NP test rejects H_0 when $S_L \ge \tau_L$ for a threshold τ_L. Under H_0, for any $s > 0$,

$$\mathbb{P}_0(S_L \ge \tau_L) = \mathbb{P}_0\big(2^{sS_L} \ge 2^{s\tau_L}\big) \;\le\; 2^{-s\tau_L}\,\mathbb{E}_0[2^{sS_L}] \quad \text{(Markov)} \tag{57}$$

$$= 2^{-s\tau_L} \prod_{t=1}^{L} \mathbb{E}_0[2^{sZ_t}] = 2^{-s\tau_L} \prod_{t=1}^{L} \sum_y P_t^{(0)}(y)\Big(\tfrac{P_t^{(1)}(y)}{P_t^{(0)}(y)}\Big)^s \tag{58}$$

$$= 2^{-s\tau_L} \prod_{t=1}^{L} \sum_y P_t^{(0)}(y)^{1-s} P_t^{(1)}(y)^s. \tag{59}$$

Taking $s = 1$ gives $\mathbb{E}_0[2^{S_L}] = 1$ and hence $\mathbb{P}_0(S_L \ge \tau_L) \le 2^{-\tau_L}$. Choosing $\tau_L = \log_2(1/\alpha)$ ensures the level constraint $\mathbb{P}_0(\text{reject } H_0) \le \alpha$.

Under H_1, for any $s \in (0,1)$,

$$\mathbb{P}_1(S_L \le \tau_L) = \mathbb{P}_1\big(2^{-sS_L} \ge 2^{-s\tau_L}\big) \;\le\; 2^{s\tau_L}\,\mathbb{E}_1[2^{-sS_L}] \quad \text{(Markov)} \tag{60}$$

$$= 2^{s\tau_L} \prod_{t=1}^{L} \mathbb{E}_1[2^{-sZ_t}] = 2^{s\tau_L} \prod_{t=1}^{L} \sum_y P_t^{(1)}(y)\Big(\tfrac{P_t^{(0)}(y)}{P_t^{(1)}(y)}\Big)^s \tag{61}$$

$$= 2^{s\tau_L} \prod_{t=1}^{L} \sum_y P_t^{(1)}(y)^{1-s} P_t^{(0)}(y)^s. \tag{62}$$

Define, in base 2, $\psi_t(s) := -\log_2 \sum_y P_t^{(1)}(y)^{1-s} P_t^{(0)}(y)^s$ and $\Psi_L(s) = \sum_{t=1}^{L} \psi_t(s)$. Then

$$\mathbb{P}_1(S_L \le \tau_L) \;\le\; 2^{\,s\tau_L - \Psi_L(s)}. \tag{63}$$

By smoothness at $s = 0$, $\psi_t(0) = 0$ and $\psi_t'(0) = D(P_t^{(1)} \| P_t^{(0)})$; moreover $\psi_t''(0)$ is the variance (in bits) of Z_t under $P_t^{(1)}$, which is finite by assumption. Hence, for s small,

$$\Psi_L(s) \;=\; s \sum_{t=1}^{L} D\big(P_t^{(1)} \| P_t^{(0)}\big) \;-\; \tfrac{1}{2}\, s^2 V_L \;+\; o(s^2 L), \quad V_L := \sum_{t=1}^{L} \mathrm{Var}_{P_t^{(1)}}(Z_t). \tag{64}$$

With $\tau_L = \log_2(1/\alpha)$ and optimizing the quadratic exponent in s yields, for all large L,

$$\mathbb{P}_1(S_L \le \tau_L) \;\le\; 2^{-\left(\sum_{t=1}^{L} D(P_t^{(1)}\|P_t^{(0)}) - \log_2(1/\alpha) - o(L)\right)}. \tag{65}$$

Therefore, given any fixed α and any β, there exists $L_0(\alpha,\beta)$ such that for all $L \ge L_0$, the miss probability is at most β whenever

$$\sum_{t=1}^{L} D\big(P_t^{(1)}\|P_t^{(0)}\big) \;\ge\; \log_2\frac{1}{\beta} \;+\; o(L), \tag{66}$$

which proves the claim. Dropping the lower-order term gives the clean sufficient rule used in the main text. □

Between Lemma 1 and Lemma 2, the picture is now clear: the watermark induces a small per-token shift from p_t to q_t whose information content is, to second order, the quadratic form of Lemma 1. Summing these local contributions across the sequence gives the total information available to the detector, and Lemma 2 translates that total into a sufficient condition for the desired power. What remains is to understand how the edit (noise) channel deforms the local shift, which is precisely the content of the next lemma.

D.2 Per-token information at $\varepsilon = 0$

For the biased family, let $I(v) = \mathbf{1}[v \in G]$ and $Z_t(\delta) = \sum_v p_t(v)e^{\delta I(v)} = (1-\gamma) + \gamma e^{\delta}$. Then

$$\log\frac{q_{t,\delta}(v)}{p_t(v)} = \delta\, I(v) - \log Z_t(\delta). \tag{67}$$

Taking expectation under $q_{t,\delta}$ and expanding at $\delta = 0$ yields (the first derivative vanishes and the second derivative equals $\mathrm{Var}_{p_t}(I) = \gamma(1-\gamma)$)

$$D\big(q_{t,\delta}\|p_t\big) = \frac{\delta^2}{2\ln 2}\,\gamma(1-\gamma) \;+\; O(\delta^3), \tag{68}$$

so in bits per token

$$D_0^{(\text{biased})} \;\approx\; \frac{\delta^2\,\gamma(1-\gamma)}{2\ln 2}. \tag{69}$$

For the bias-free family, write $R_E(v) = 1 + \Delta_E(v)$ with $\mathbb{E}_E[\Delta_E(v)] = 0$ and $\|\Delta_E\|_\infty$ small. Then

$$D\big(q_{t,E}\|p_t\big) = \sum_v p_t(v)\,(1+\Delta_E(v))\log\big(1+\Delta_E(v)\big). \tag{70}$$

Using $\log(1+x) = x - \frac{x^2}{2} + O(x^3)$ and averaging over E,

$$\mathbb{E}_E\big[D(q_{t,E}\|p_t)\big] = \frac{1}{2\ln 2}\sum_v p_t(v)\,\mathbb{E}_E[\Delta_E(v)^2] \;+\; O\left(\sum_v p_t(v)\,\mathbb{E}[|\Delta_E(v)|^3]\right). \tag{71}$$

With $\sigma^2(v) = \mathrm{Var}_E[R_E(v)]$ and $\hat\sigma^2 = \sum_v p_t(v)\sigma^2(v)$ this gives, in bits/token,

$$D_0^{(\text{bias-free})} \;\approx\; \frac{\hat\sigma^2}{2\ln 2}. \tag{72}$$

These two expressions are exactly the D_0 quantities used in the theorem.

D.3 Edits contract the signal quadratically

We now show that the edit channel scales the local perturbation by $(1-\varepsilon)$ and hence the local KL by $(1-\varepsilon)^2$ to second order. This is the key structural fact that produces the quadratic decay with the edit rate.

Lemma 3 (Local $(1-\varepsilon)^2$ contraction)**.** *Fix p on Σ and write $q = p + r$ with $\sum_v r(v) = 0$ and $\|r\|_\infty$ small. Let $p_\varepsilon = T_\varepsilon(p) = (1-\varepsilon)p + \varepsilon U$ and $q_\varepsilon = T_\varepsilon(q) = (1-\varepsilon)q + \varepsilon U$. Then*

$$D\big(q_\varepsilon \| p_\varepsilon\big) \;=\; (1-\varepsilon)^2 \,\frac{1}{2\ln 2} \sum_v \frac{r(v)^2}{p_\varepsilon(v)} \;+\; o\!\left(\sum_v \frac{r(v)^2}{p(v)}\right). \tag{73}$$

In particular, when p_ε and p are boundedly comparable (which holds for every fixed $\varepsilon > 0$), we have

$$D\big(q_\varepsilon \| p_\varepsilon\big) \;=\; (1 + o(1))\,(1-\varepsilon)^2\, D(q\|p). \tag{74}$$

Proof. Since $q_\varepsilon - p_\varepsilon = (1-\varepsilon)r$, apply Lemma 1 at the reference p_ε:

$$D(q_\varepsilon \| p_\varepsilon) = \frac{1}{2\ln 2} \sum_v \frac{\big((1-\varepsilon)r(v)\big)^2}{p_\varepsilon(v)} + R_\varepsilon \tag{75}$$

$$= (1-\varepsilon)^2 \cdot \frac{1}{2\ln 2} \sum_v \frac{r(v)^2}{p_\varepsilon(v)} + R_\varepsilon, \tag{76}$$

with $R_\varepsilon = o\big(\sum_v r(v)^2/p(v)\big)$ as $\|r\|_\infty \to 0$. The comparability $p_\varepsilon(v) \in [(1-\varepsilon)p(v),\ (1-\varepsilon)p(v) + \varepsilon/|\Sigma|]$ yields the stated equivalence. □

D.4 From per-token information to sequence-level reliability

Let $\{p_t\}_{t=1}^L$ and $\{q_t\}_{t=1}^L$ denote the baseline and watermarked conditionals, respectively. Under the edit channel we observe $\{p_{t,\varepsilon}\}$ and $\{q_{t,\varepsilon}\}$. The KL chain rule aggregates local contributions along the sequence and shows that conditioning can only reduce KL on average; thus the unconditional sum of per-token KLs is a valid (and often tight) proxy for the total.

Lemma 4 (Additivity bound for total information)**.** *For the binary test $H_0 : \prod_t p_{t,\varepsilon}$ versus $H_1 : \prod_t q_{t,\varepsilon}$, the total KL satisfies*

$$D\left(\prod_{t=1}^L q_{t,\varepsilon} \;\Big\|\; \prod_{t=1}^L p_{t,\varepsilon}\right) = \sum_{t=1}^L \mathbb{E}_{H_1}\Big[D\big(q_{t,\varepsilon}(\cdot \mid Y_{<t}) \| p_{t,\varepsilon}(\cdot \mid Y_{<t})\big)\Big] \tag{77}$$

$$\leq \sum_{t=1}^L D(q_{t,\varepsilon}\|p_{t,\varepsilon}). \tag{78}$$

If the embedder is memoryless and per-step statistics are homogeneous, the equality reduces to the sum of identical per-token KLs.

Proof. The equality is the KL chain rule. The inequality is Jensen's inequality: averaging over histories (conditioning) cannot increase KL. □

Combining Lemma 3 with Lemma 4 yields the total information available to the detector,

$$C(\varepsilon) \;:=\; \sum_{t=1}^L D(q_{t,\varepsilon}\|p_{t,\varepsilon}) \;\approx\; L\,(1-\varepsilon)^2\, D_0, \tag{79}$$

with D_0 given by equation 69 or equation 72.

D.5 Power condition and the "knee" edit rate

We now translate total information into a sufficient condition for the target power. Applying Lemma 2 with total signal $C(\varepsilon)$ gives

$$L\,(1-\varepsilon)^2\, D_0 \;\geq\; \log_2 \frac{1}{\beta}, \tag{80}$$

which guarantees miss probability at most β. Solving for ε produces the *knee*—the maximal edit rate compatible with the target power:

$$\varepsilon_\beta(L, D_0) \;=\; 1 - \sqrt{\frac{\log_2(1/\beta)}{L\, D_0}}. \tag{81}$$

This completes the proof of Theorem 2 once the family-specific expressions for D_0 from equation 69 and equation 72 are substituted.

D.6 Impossibility region and qualitative behavior

The impossibility region follows immediately: whenever the total information falls below the required threshold, no level-α detector can meet the target power.

Proposition 1 (Impossibility region)**.** *For fixed (L, β) and per-token information D_0, if*

$$L\,(1-\varepsilon)^2\, D_0 \;<\; \log_2 \frac{1}{\beta}, \tag{82}$$

then detection at power $1-\beta$ is impossible. Equivalently, no method can succeed for $\varepsilon > \varepsilon_\beta(L, D_0)$.

Proof. This is the contrapositive of Lemma 2 applied to the total sequence divergence. □

In the small-signal regime with independent contributions, the separation of likelihood-ratio scores under H_0 and H_1 is governed by the same total KL and therefore by $L(1-\varepsilon)^2 D_0$. Once this quantity drops below the threshold $\log_2(1/\beta)$, the score distributions are no longer reliably separable and operating characteristics converge to chance.

D.7 Assumptions, approximations, and scope of validity

The derivation operates in a small-signal regime. For the biased family this means $|\delta| \ll 1$; for the bias-free family it means $\|\Delta_E\|_\infty \ll 1$ and $p_t(v)$ bounded away from zero. Lemma 1 quantifies the approximation error and shows it is lower order relative to the quadratic term in the perturbation. The $(1-\varepsilon)^2$ contraction in Lemma 3 is a local statement around the operating point and uses the quadratic form that defines the local KL (equivalently, Fisher information). The aggregation argument uses the KL chain rule; for memoryless embedding with homogeneous per-step statistics, the total KL is exactly the sum of per-step KLs, whereas in general it is upper bounded by that sum, which suffices for a *sufficient* power condition. Lastly, Lemma 2 is invoked as a sufficiency result: for independent per-token contributions with regularity, the type-II error exponent matches the KL (Chernoff–Stein achievability), and the base-2 normalization cleanly produces the threshold $\log_2(1/\beta)$ in bits.

D.8 Worked numeric examples

For illustration, take $L = 1000$ and power targets $1-\beta \in \{0.90, 0.95, 0.99\}$, so that

$$\log_2(1/\beta) \in \{3.322,\; 4.322,\; 6.644\}. \tag{83}$$

If the total noise-free information is $LD_0 = 10$ bits (e.g., $D_0 = 0.01$ bits/token), the knees are

$$\varepsilon_{90} \approx 0.424, \qquad \varepsilon_{95} \approx 0.343, \qquad \varepsilon_{99} \approx 0.185. \tag{84}$$

For the biased family with $\gamma = 0.25$, achieving $D_0 = 0.01$ requires approximately

$$\delta \;\approx\; \sqrt{\frac{2\ln 2\, D_0}{\gamma(1-\gamma)}} \;\approx\; 0.27, \tag{85}$$

while for the bias-free family one needs $\hat{\sigma}^2 \approx 2\ln 2\, D_0 \approx 0.0139$.

D.9 Conclusion of the proof

Combining (i) the small-signal per-token KL for the biased and bias-free families, (ii) the quadratic attenuation $(1-\varepsilon)^2$ under the edit channel, (iii) the chain rule aggregation across L positions, and (iv) Stein's sufficient condition for miss probability β, yields the theorem's sufficiency condition

$$L(1-\varepsilon)^2 D_0 \;\geq\; \log_2(1/\beta), \tag{86}$$

and the corresponding knee

$$\varepsilon_\beta(L, D_0) = 1 - \sqrt{\frac{\log_2(1/\beta)}{L D_0}}. \tag{87}$$

The impossibility region and qualitative behavior beyond the knee discussed in the main text follow directly. □

Proof of Corollary 1

The corollary merges the baseline operating boundary with its stealth-aware tightening. For the baseline part, Theorem 2 asserts that reliable detection at power $1-\beta$ requires $L(1-\varepsilon)^2 D_0 \geq \log_2(1/\beta)$. Therefore, for any $\varepsilon > \varepsilon_\beta(L, D_0)$ with ε_β as defined above, the inequality is violated and reliable detection is unattainable.

For the stealth-aware part, suppose an outsider may pool M tokens and we require that the watermarked and baseline distributions remain within total variation τ on that pooled sample. Pinsker's inequality, together with the base conversion from nats to bits, implies the per-token information constraint $D_0 \leq (2/\ln 2)\,\tau^2/M$. Substituting this into the baseline condition yields

$$L(1-\varepsilon)^2 \frac{2\tau^2}{M \ln 2} \;\geq\; \log_2(1/\beta) \quad\Longrightarrow\quad \varepsilon \;\leq\; 1 - \sqrt{\frac{\log_2(1/\beta)}{L} \cdot \frac{M \ln 2}{2\tau^2}}. \tag{88}$$

Thus any edit rate exceeding the right-hand side is infeasible under the stated stealth constraint. □

E Proof of Theorem 3

All logarithms are base 2, so every divergence and information quantity is measured in bits. The proof is organized into several stages, each of which builds toward the statement of the theorem. We begin with the local information contributed per token by biased and bias-free watermarking families. We then quantify the attenuation introduced by the substitution edit channel and extend this to sequences using the KL chain rule. We next invoke the Chernoff–Stein lemma to obtain a sufficiency condition for reliable detection. After this, we translate stealth requirements into information caps using Pinsker's inequality. Finally, we combine these pieces into the composite loss, which determines the optimal operating point, and analyze how the allocation between families should be made. The proof concludes by identifying conditions under which distribution-preserving watermarking strictly dominates.

E.1 Per-token information in the small-signal regime

We begin with the biased (tilt) family. At a given position with baseline conditional distribution p_t over the vocabulary Σ, a key-selected subset $G \subseteq \Sigma$ with baseline mass $\gamma = \sum_{v\in G} p_t(v)$ is exponentially tilted with parameter $\delta \in \mathbb{R}$. This produces the conditional

$$q_{t,\delta}(v) \;=\; \frac{p_t(v)\, e^{\delta\, \mathbf{1}[v\in G]}}{Z_t(\delta)}, \qquad Z_t(\delta) = (1-\gamma) + \gamma e^{\delta}. \tag{89}$$

Expanding $\log_2(q_{t,\delta}(v)/p_t(v))$ around $\delta = 0$ and retaining the leading nonzero term gives

$$D\big(q_{t,\delta} \,\|\, p_t\big) = \frac{\gamma(1-\gamma)}{2\ln 2}\,\delta^2 + O(\delta^3). \tag{90}$$

Thus, the small-signal per-token information is

$$D_0^{\mathrm{B}} \;\approx\; \frac{\gamma(1-\gamma)}{2\ln 2}\,\delta^2, \tag{91}$$

which is maximized at $\gamma^\star = \frac{1}{2}$ for fixed D_0.

For the bias-free family, the watermarked conditional is a mean-one reweighting $q_{t,E}(v) = p_t(v)R_E(v)$ with $\mathbb{E}[R_E(v)] = 1$. Writing $R_E(v) = 1 + \Delta_E(v)$ and expanding $\log(1 + \Delta_E(v))$ shows that the quadratic variance term dominates, yielding

$$D_0^{\mathrm{BF}} \;\approx\; \frac{\hat{\sigma}^2}{2\ln 2}, \qquad \hat{\sigma}^2 = \sum_v p_t(v)\,\mathrm{Var}[R_E(v)]. \tag{92}$$

E.2 Attenuation under edits

Each token passes through the substitution channel

$$T_\varepsilon(P) \;:=\; (1-\varepsilon)P + \varepsilon U, \tag{93}$$

where U is uniform on Σ. If $q = p + r$ with $\sum_v r(v) = 0$, then $T_\varepsilon(q) - T_\varepsilon(p) = (1-\varepsilon)r$. Since KL divergence is locally quadratic in r, the attenuation factor is squared, giving

$$D\big(T_\varepsilon(q)\,\|\,T_\varepsilon(p)\big) = (1-\varepsilon)^2\,D(q\|p)\,(1+o(1)). \tag{94}$$

Consequently, for either family the per-token information after edits is

$$D_\varepsilon \;\approx\; (1-\varepsilon)^2 D_0. \tag{95}$$

E.3 Sequence-level information accumulation

The KL chain rule extends the per-token information to sequences. Writing $p_{t,\varepsilon} = T_\varepsilon(p_t)$ and $q_{t,\varepsilon} = T_\varepsilon(q_t)$, one obtains

$$D\left(\prod_{t=1}^{T} q_{t,\varepsilon}\,\middle\|\,\prod_{t=1}^{T} p_{t,\varepsilon}\right) \le \sum_{t=1}^{T} D(q_{t,\varepsilon}\|p_{t,\varepsilon}). \tag{96}$$

In the homogeneous small-signal regime each summand is approximately D_ε, so the total usable signal is

$$C(\varepsilon) \;\approx\; T(1-\varepsilon)^2 D_0. \tag{97}$$

E.4 Reliability requirement via Chernoff–Stein

A level-α Neyman–Pearson test achieves miss probability at most β if the sequence-level KL under the alternative exceeds $\log_2(1/\beta)$. Combining this condition with equation 97 gives

$$D_0 \;\ge\; D_{\mathrm{req}}(\varepsilon, T, \beta) \;:=\; \frac{\log_2(1/\beta)}{T(1-\varepsilon)^2}. \tag{98}$$

This inequality captures the robustness requirement: a minimum information budget per token is needed to guarantee detection.

E.5 Stealth constraints via Pinsker

Pinsker's inequality in nats yields $\mathrm{TV} \le \sqrt{D_{\mathrm{nat}}/2}$, and converting bits to nats gives $D_{\mathrm{nat}} = M(\ln 2)D_0$ for M pooled tokens. Thus,

$$\mathrm{TV} \;\le\; \sqrt{\tfrac{\ln 2}{2}\,M\,D_0}. \tag{99}$$

Imposing a budget $\mathrm{TV} \le \tau$ leads to the stealth cap

$$D_0 \;\le\; D_{\mathrm{stealth}}(M, \tau) \;:=\; \frac{2\tau^2}{M\ln 2}. \tag{100}$$

E.6 Minimization of the composite loss

The composite loss is

$$\mathcal{L}(\theta;\varepsilon,M,\tau) = \lambda_r[\log_2(1/\beta) - T(1-\varepsilon)^2 D_0(\theta)]_+ + \lambda_q \,\mathrm{TV}_{\mathrm{pen}}(D_0(\theta);M) + \lambda_a \,\mathrm{Amp}(\theta). \quad (101)$$

Because the hinge vanishes once D_0 reaches the required threshold, while both detectability and amplitude penalties increase with D_0, the optimizer must select the smallest feasible D_0. This gives

$$D^\star = \min\{D_{\mathrm{stealth}}(M,\tau),\, D_{\mathrm{BF}}^{\max} + D_{\mathrm{B}}^{\max}\}, \qquad D^\star \;\geq\; D_{\mathrm{req}}(\varepsilon,T,\beta). \quad (102)$$

If this inequality cannot be satisfied, reliable detection is impossible at the given edit rate.

E.7 Optimal allocation between families

With $D^\star$ fixed, the TV penalty depends only on its value, not on the split between families. Hence the allocation minimizes the amplitude term. Since

$$\hat{\sigma}^2 = 2\ln 2\, D_0^{\mathrm{BF}}, \qquad \delta^2 = 8\ln 2\, D_0^{\mathrm{B}} \quad (\gamma = \tfrac{1}{2}), \quad (103)$$

the amplitude penalty is

$$\lambda_a\left(\sqrt{2\ln 2}\sqrt{D_0^{\mathrm{BF}}} + \sqrt{8\ln 2}\sqrt{D_0^{\mathrm{B}}}\right). \quad (104)$$

This is minimized by maximizing the allocation to BF, subject to its budget. Therefore,

$$D_0^{\mathrm{BF}\star} = \min\{D^\star,\, D_{\mathrm{BF}}^{\max}\}, \qquad D_0^{\mathrm{B}\star} = D^\star - D_0^{\mathrm{BF}\star}. \quad (105)$$

The corresponding parameter values are

$$\hat{\sigma}^{2\star} = 2\ln 2\, D_0^{\mathrm{BF}\star}, \qquad \delta^\star = \sqrt{8\ln 2\, D_0^{\mathrm{B}\star}}, \qquad \gamma^\star = \tfrac{1}{2}. \quad (106)$$

If $D_{\mathrm{req}}(\varepsilon,T,\beta) \leq D_{\mathrm{BF}}^{\max}$, the optimizer chooses pure BF; otherwise BF is saturated and the remainder is realized with B.

E.8 Dominance of distribution-preserving watermarking

Finally, we examine when distribution-preserving watermarking is preferable. Suppose K positions are marked and the verifier corrects up to t errors. If $X \sim \mathrm{Binomial}(K, 1-\varepsilon)$ counts surviving marks, then

$$\Pr[X < K - t] \;\leq\; \exp\left(-2K\Big((1-\varepsilon) - t/K\Big)^2\right). \quad (107)$$

Thus DP achieves miss probability at most β whenever

$$(1-\varepsilon) \;\geq\; \frac{t}{K} + \sqrt{\frac{\ln(1/\beta)}{2K}}. \quad (108)$$

Because DP leaves the token distribution unchanged, it yields zero detectability and, therefore, strictly dominates any statistical scheme meeting the same robustness target. In this region, DP is optimal; outside of it, the statistical allocation of equation 105 applies.

E.9 Conclusion

Combining the small-signal identities equation 91–equation 95, the sequence accumulation equation 97, the reliability requirement equation 98, the stealth cap equation 100, the composite loss equation 101, the allocation rule equation 105, and the DP dominance condition equation 108 establishes the full structure of the hybrid watermarking strategy and completes the proof of Theorem 3.

Table 1: Performance evaluation of Biased (KGW Kirchenbauer et al. (2023), Unigram Zhao et al. (2023)), Bias-free (DiPMark Wu et al. (2024), HCW Hu et al. (2024)), and undetectable CGW Christ et al. (2024) watermarking schemes on Llama-2-7B and Mistral-7B. For all cases, we evaluate robustness metrics (Reliable detection with key in presence of noise): AUROC, TPR at 1% FPR, and F1 at 1% FPR. We also evaluate detectability metrics (detection without key using statistical tests) via p-score and z-score.

Model	Attack	Method	Robustness (with key)			Detectability (no key)	
			AUROC	TPR@1%	F1@1%	p-score[a]	z-score[b]
Llama-2-7B	Reference (no paraphrasing)	KGW (Biased)	0.99	1.000	0.995	0.72	30.1
		Unigram (Biased)	0.99	1.000	0.995	0.68	11.2
		DiPMark (Bias-free)	0.99	1.000	0.995	0.31	43.2
		HCW (Bias-free)	0.99	1.000	0.995	0.28	105.1
		CGW (Dist-pres.)	0.99	1.000	0.995	—	-5.8
		Optimal Hybrid[c]	0.99	1.000	0.995	—	-7.8
	DIPPER (avg $\epsilon = 0.25$)	KGW (Biased)	0.860	0.640	0.780	0.72	9.6
		Unigram (Biased)	0.875	0.665	0.795	0.68	8.8
		DiPMark (Bias-free)	0.895	0.800	0.865	0.31	3.9
		HCW (Bias-free)	0.905	0.820	0.875	0.28	3.4
		CGW (Dist-pres.)	0.500	0.150	0.230	—	-10.2
		Optimal Hybrid[c]	0.910	0.835	0.885	—	5.7
	OPT-2.7B (avg $\epsilon = 0.15$)	KGW (Biased)	0.780	0.590	0.720	0.72	8.4
		Unigram (Biased)	0.790	0.615	0.740	0.68	7.9
		DiPMark (Bias-free)	0.905	0.855	0.900	0.31	3.6
		HCW (Bias-free)	0.920	0.880	0.915	0.28	3.1
		CGW (Dist-pres.)	0.502	0.310	0.420	—	-5.4
		Optimal Hybrid[c]	0.930	0.895	0.922	—	4.5
Mistral-7B	Reference (no paraphrasing)	KGW (Biased)	0.99	1.000	0.995	0.69	27.8
		Unigram (Biased)	0.99	1.000	0.995	0.66	10.5
		DiPMark (Bias-free)	0.99	1.000	0.995	0.34	39.5
		HCW (Bias-free)	0.99	1.000	0.995	0.26	98.7
		CGW (Dist-pres.)	0.99	1.000	0.995	—	-12.5
		Optimal Hybrid[c]	0.99	1.000	0.995	—	-11.0
	DIPPER (avg $\epsilon = 0.25$)	KGW (Biased)	0.845	0.615	0.765	0.71	9.0
		Unigram (Biased)	0.860	0.640	0.780	0.67	8.2
		DiPMark (Bias-free)	0.885	0.785	0.860	0.33	4.1
		HCW (Bias-free)	0.895	0.805	0.872	0.29	3.5
		CGW (Dist-pres.)	0.500	0.135	0.210	—	-8.9
		Optimal Hybrid[c]	0.902	0.820	0.880	—	7.6
	OPT-2.7B (avg $\epsilon = 0.15$)	KGW (Biased)	0.760	0.565	0.705	0.71	8.2
		Unigram (Biased)	0.770	0.585	0.720	0.67	7.7
		DiPMark (Bias-free)	0.890	0.840	0.890	0.32	3.8
		HCW (Bias-free)	0.910	0.865	0.902	0.29	3.2
		CGW (Dist-pres.)	0.501	0.285	0.400	—	-9.7
		Optimal Hybrid[c]	0.922	0.875	0.910	—	8.8

[a] p-score detectability metric reported by Gloaguen et al. (2025), which is watermark specific, hence left blank for CGW Christ et al. (2024) and proposed optimal hybrid watermarking scheme.
[b] z-score detectability metric reported by Liu et al. (2025), with negative score meaning less detectability.
[c] Proposed Pareto-optimal hybrid watermarking scheme by Theorem 3.

F Additional Experimental Results and Discussion

For each base model in Table 1 (Llama 2 7B and Mistral 7B), we evaluate three editing conditions and then measure detection strength and third-party detectability for each watermarking scheme. The two paraphrasing conditions apply DIPPER Krishna et al. (2023) with a token editing rate of $\epsilon = 0.25$ and OPT 2.7B, prompted with "Rewrite the following paragraph:" with an average $\epsilon = 0.15$, which induces higher and lower token changes, respectively. For every condition, we report detection metrics with access to the key (area under the ROC curve, TPR at 1% FPR, and F1 at 1% FPR) and detectability metrics without the key using p-score and z-score from black box statistical tests Gloaguen et al. (2025); Liu et al. (2025).

We evaluate the following families and instances: Biased (KGW Kirchenbauer et al. (2023), Unigram Zhao et al. (2023)), Bias free (DiPMark Wu et al. (2024), HCW Hu et al. (2024)), and distribution preserving CGW Christ et al. (2024), along with our Optimal Hybrid (Theorem 3). This setup

places each watermarking scheme at a point on the plane that balances detection strength against detectability, revealing how that point moves as edit intensity changes under using DIPPER and OPT 2.7B paraphrasing attacks.

Across both models, the detection–detectability tradeoff primarily depends on the watermarking family, rather than the underlying LLM. In the no-paraphrasing condition (reference), all methods achieve near-perfect detection strength; however, detectability differs markedly: CGW sits near the low detectability corner, KGW and Unigram are easily flagged statistically, and DiPMark and HCW occupy the middle. Under DIPPER with average $\epsilon = 0.25$, CGW loses most of its detection strength, DiPMark and HCW maintain midrange values, and KGW and Unigram lie between these extremes; OPT 2.7B paraphrasing with average $\epsilon = 0.15$ causes a milder shift but preserves the same ordering. A single fixed family does not satisfy both needs over the full range of edits. In contrast, the Optimal Hybrid uses a simple estimate of edit intensity to select the active family, moving toward CGW when edits are light to keep detectability low and shifting toward HCW or KGW/Unigram as edits increase to keep high TPR at a fixed false positive rate. The empirical results align with our theory, and the closely matched trends for Llama and Mistral indicate that placement on the accuracy–detectability plane is driven by the watermarking type rather than the model type.

Extension to Training-Time Watermarks. While our analysis focuses on inference-time watermarking where the sampling distribution is modified during text generation, recent work in Gu et al. (2024) has explored embedding watermarks directly into model parameters during training. These training-time approaches learn weights-based watermarking through distillation, enabling models to naturally generate watermarked text under standard decoding algorithms without specialized sampling procedures. Our information-theoretic framework provides insights into these methods: since detectability depends solely on the sampling distribution rather than model architecture (Theorem 1), training-time watermarks must fundamentally alter the model's learned distribution p_θ to approximate the watermarked distribution q. This introduces additional challenges: the watermark signal becomes vulnerable to fine-tuning attacks that can remove the embedded patterns, and the distillation process itself incurs a quality-detectability penalty beyond our theoretical bounds. Nevertheless, training-time watermarks offer practical advantages for open-source models where users control the decoding process, suggesting that hybrid approaches combining training-time embedding with inference-time enhancement may achieve better robustness-detectability trade-offs than either method alone.

G WATERMARKING AS COVERT CHANNELS

Modern watermark detectors aggregate a small per-token statistical signal and then apply a Neyman-Pearson test to distinguish watermarked text from baseline text. The same statistical signal can be intentionally controlled to convey side information, thereby turning watermarking mechanisms into covert channels. Let D_0 denote the noise-free per-token information in bits per token contributed by a given watermark family. When the text passes through a substitution edit channel with edit rate ε, this quantity contracts quadratically:

$$D_\varepsilon \approx (1-\varepsilon)^2 D_0. \tag{109}$$

For a sequence of length L, the total detector signal available after edits is

$$C(\varepsilon) \approx L(1-\varepsilon)^2 D_0. \tag{110}$$

In the biased green list family, where a subset of the vocabulary of baseline mass γ is exponentially tilted by a factor δ, a second-order expansion gives

$$D_0 \approx \frac{\delta^2 \gamma(1-\gamma)}{2\ln 2}. \tag{111}$$

In the bias-free family, where unbiased reweighting with variance $\hat{\sigma}^2$ is applied, the corresponding expression is

$$D_0 \approx \frac{\hat{\sigma}^2}{2\ln 2}. \tag{112}$$

A level α Neyman Pearson test achieves miss probability at most β whenever

$$L(1-\varepsilon)^2 D_0 \;\geq\; \log_2\Big(\frac{1}{\beta}\Big). \tag{113}$$

Table 2: Covert use of common watermark families (width safe TabularX)

Family	How Alice encodes	How Bob decodes
Greedy tie breaking	Use a pre-shared set of tie-breaking keys. Select k_m to represent message m and add a small term $\epsilon\,\mathrm{PRF}_{k_m}(t, v)$ to logits when two top scores fall within a narrow band.	Given the prompt and model, simulate greedy decisions under each $k \in \mathcal{K}$ and choose the key that reproduces the observed tie outcomes.
Biased green list (tilt)	Partition m into w bit chunks and use each chunk to seed the green list in its window. Apply the exponential tilt $q_t \propto p_t \exp\{\delta\,\mathbf{1}[v \in G_t]\}$ with small δ.	Regenerate the candidate green lists and evaluate the matched filter score of green token counts per window to select the chunk.
Bias-free (variance)	Use keyed permutations or codes R_{E_t} with $\mathbb{E}[R_{E_t}(v)] = 1$ so that $\mathbb{E}[q_t] = p_t$ while the variance carries the information.	Apply the family-specific key verified statistic and perform maximum likelihood over the 2^w codes in each window.
Distribution preserving (PRF-driven RNG)	Replace $U_t \sim \mathrm{Uniform}[0, 1]$ by $U_t = \mathrm{PRF}_{k,m}(\mathrm{context}_t, t)$ and draw $v_t = F_{p_t}^{-1}(U_t)$, leaving one step marginals unchanged.	Resimulate the RNG stream with k and candidate $\hat{m}$ and select the $\hat{m}$ that reproduces the observed sequence.

Table 3: Message size scaling on length L and edits at rate ε (width safe TabularX)

Family	Message size on length L	Outsider evidence driver
Greedy tie breaking	$\log_2 \lvert\mathcal{K}\rvert$ once per document	Large sequence drift relative to the stochastic baseline.
Biased green list (tilt)	$\Theta(\sqrt{L})$ under a fixed outsider mixture budget (Lemma 5)	$D_0 \approx \delta^2\gamma(1-\gamma)/(2\ln 2)$ and edited signal $L(1-\varepsilon)^2 D_0$.
Bias free (variance)	$\Theta(\sqrt{L})$ under a fixed outsider mixture budget (Lemma 5)	$D_0 \approx \hat{\sigma}^2/(2\ln 2)$ and edited signal $L(1-\varepsilon)^2 D_0$.
Distribution preserving (PRF RNG)	$\Theta(L)$ in a single pass; repeated queries reveal determinism unless the seed is ephemeral	One step marginals match the baseline; a single pass outsider sees no local drift, but identical replays can expose determinism.

Solving for the maximum admissible edit rate that still guarantees power $1 - \beta$ yields

$$\varepsilon_\beta(L, D_0) \;=\; 1 \;-\; \sqrt{\frac{\log_2(1/\beta)}{L D_0}}. \tag{114}$$

This expression shows that there is no universal critical edit rate; instead, performance depends jointly on L, D_0, and β.

G.1 Turning watermark rules into channels

Alice and Bob share a secret key k. During generation, Alice steers a standard probability-modifying watermark family to encode a message, and Bob decodes it using the matched key and verified statistics. An outsider observes only the text and is unaware of k. The constructions below are representative and capture the essential scaling laws. The receiver always applies the detector that is matched to the family and keyed to k.

G.2 Capacity versus detectability: a square root law

The following lemma (based on Theorem 2) formalizes the relationship between achievable message size and outsider evidence. It captures the square root scaling for biased and bias-free families under a realistic stealth requirement on the outsider mixture, and it clarifies the stronger constraint that arises if one demands small drift for every message separately.

Lemma 5 (Capacity detectability law for watermark driven channels)**.** *Let a watermark family contribute D_0 bits of information per token. A covert transmitter chooses a message $W \in \{1, \dots, M\}$ uniformly and uses a secret key so that the outsider observes the mixture $Q = \frac{1}{M}\sum_{w=1}^{M} Q_w$. Then:*

(a) Mixture budget. *If the outsider mixture satisfies $D(Q\|P) \le C_\star$ for a constant $C_\star$ independent of L, then for biased and bias free families*

$$\log M \;=\; \Theta(\sqrt{L}) \tag{115}$$

in the noise-free case, and

$$\log M \;=\; \Theta\big((1-\varepsilon)^2\sqrt{L}\big) \tag{116}$$

under the substitution edit channel at rate ε.

(b) Per message pooling. *If one imposes the stronger constraint $\mathrm{TV}(Q_w, P) \le \tau$ for every message w, then Pinsker's inequality gives $D_0 \le 2\tau^2/(L\ln 2)$ and hence*

$$\log M \;=\; O(1). \tag{117}$$

(c) Linear growth requires vanishing per token drift. *Any scheme that achieves $\log M = \omega(\sqrt{L})$ while keeping $D(Q\|P) \le C_\star$ must satisfy $D_0 \to 0$ at the one-step margin, that is, it must be distribution preserving.*

G.3 Proof of Lemma 5

We first recall the small signal identities that underlie all bounds. For the biased family,

$$D_0 \;\approx\; \frac{\delta^2\,\gamma(1-\gamma)}{2\ln 2}\,, \tag{118}$$

and for the bias-free family,

$$D_0 \;\approx\; \frac{\hat{\sigma}^2}{2\ln 2}\,. \tag{119}$$

Under the substitution channel, the per-token information contracts as

$$D_\varepsilon \;\approx\; (1-\varepsilon)^2 D_0\,, \tag{120}$$

so the total sequence level signal equals

$$C(\varepsilon) \;\approx\; L(1-\varepsilon)^2 D_0\,. \tag{121}$$

A level α Neyman Pearson test reaches miss probability at most β once

$$L(1-\varepsilon)^2 D_0 \;\ge\; \log_2\Big(\frac{1}{\beta}\Big). \tag{122}$$

Achievability under the mixture constraint. Consider a sparse activity design. Fix $\theta_L = c/\sqrt{L}$ with $c > 0$. Using the secret key, mark each position active independently with probability θ_L; inactive positions are sampled from the baseline. On active positions, apply a constant tilt $\delta = \delta_0$ and select the green list using successive message chunks. The outsider mixture at a given token is

$$(1-\theta_L)p_t + \theta_L q_{t,\delta_0}\,, \tag{123}$$

and a second-order expansion gives its KL to p_t as

$$D((1-\theta_L)p_t + \theta_L q_{t,\delta_0} \,\|\, p_t) = \frac{\theta_L^2}{2\ln 2}\sum_v \frac{\big(q_{t,\delta_0}(v) - p_t(v)\big)^2}{p_t(v)} + o(\theta_L^2) \tag{124}$$

$$\approx \theta_L^2 \cdot \frac{\delta_0^2\,\gamma_t(1-\gamma_t)}{2\ln 2}\,. \tag{125}$$

Summing over L tokens yields

$$D(Q\|P) \;\approx\; L\,\theta_L^2 \cdot \frac{\delta_0^2\,\overline{\gamma(1-\gamma)}}{2\ln 2} \;=\; O(1)\,, \tag{126}$$

since $L\theta_L^2 = c^2$ and δ_0 is constant. Thus the mixture divergence remains bounded uniformly in L. Conditioned on the key there are $T = \theta_L L = c\sqrt{L}$ active positions. On each active position, the matched statistic provides a constant positive information increment $\kappa > 0$. Standard concentration for log likelihood ratios then gives reliable decoding, provided

$$\log M \;\le\; \frac{1}{2}\,T\,\kappa \;-\; \omega(1) \;=\; \Theta(\sqrt{L})\,. \tag{127}$$

Under edits at rate ε, each active increment contracts by $(1-\varepsilon)^2$, so the same argument yields

$$\log M \;=\; \Theta\big((1-\varepsilon)^2\sqrt{L}\big). \tag{128}$$

Converse under the mixture constraint. Let Q_w denote the distribution induced by message w and $Q = \frac{1}{M}\sum_w Q_w$ the outsider mixture. The mutual information satisfies

$$I(W; Y_{1:L}) \;=\; \frac{1}{M}\sum_{w=1}^{M} D(Q_w\|Q)\,. \tag{129}$$

The log sum inequality together with the small signal expansion that controls $D(Q\|P)$ implies that the average squared perturbation around P is of order $1/L$, which limits the aggregate distinguishability across messages to order $\sqrt{L}$. A sphere packing bound for multi-hypothesis testing with total information budget of order $\sqrt{L}$, therefore, yields

$$\log M \;\le\; c_1\sqrt{L} + O(1)\,, \tag{130}$$

for a constant c_1 determined by the family and the map from watermark strength to D_0. The same $(1-\varepsilon)^2$ contraction applies under edits.

Per message pooling constraint. If for every w one requires $\mathrm{TV}(Q_w, P) \le \tau$, then Pinsker and unit conversion imply

$$LD_0 \;\le\; \frac{2\tau^2}{\ln 2}\,, \tag{131}$$

so $D_0 = O(1/L)$ and any two messages have only a constant order separation across the entire text. Reliable decoding is then possible for at most a constant number of hypotheses, which proves the stated order.

Distribution preserving case. If $D_0 = 0$ at the one-step margin, for example, by replacing the RNG with a pseudorandom stream, then one can place one bit of seed-controlled entropy per token without changing one-step marginals. In a single-pass setting, this allows

$$\log M \;=\; \Theta(L)\,, \tag{132}$$

although repeated queries with the same seed reveal determinism unless the seed is refreshed, making this a pure covert channel rather than a forensic watermark.

G.4 Reliability under edits: knee and AUROC

With total usable information

$$C(\varepsilon) \;\approx\; L(1-\varepsilon)^2 D_0\,, \tag{133}$$

the sufficiency condition for miss probability β is

$$L(1-\varepsilon)^2 D_0 \;\ge\; \log_2\Big(\frac{1}{\beta}\Big), \tag{134}$$

and the corresponding knee is

$$\varepsilon_\beta(L, D_0) \;=\; 1 \;-\; \sqrt{\frac{\log_2(1/\beta)}{LD_0}}\,. \tag{135}$$

Beyond this point, the score distributions of the likelihood ratio test largely overlap, and the area under the ROC curve approaches 0.5 with only finite sample fluctuations.