Catch-22: On the Fundamental Tradeoff Between Detectability and Robustness in LLM Watermarking

Anonymous Authors[1]

Abstract

Large language models generate text through probabilistic token sampling, a mechanism increasingly leveraged for inference-time watermarking to verify AI-generated content. We present an information-theoretic framework that characterizes the trade-off between robustness to text editing and detectability by keyless observers, where detectability bounds are information-theoretic and computational attainability depends on detector access. Central to our analysis is an additive, usable Kullback-Leibler (KL) information budget that governs hypothesis testing separability between watermarked and unwatermarked outputs subject to a stealth constraint. This budget induces a hierarchy of detectability across watermark families: distribution-preserving schemes exhibit zero statistical drift, while probability-modifying schemes at both token and sentence levels accumulate detectable signal with sequence length. When text editing is modeled as a noise process, the usable KL budget contracts quadratically with edit rate for token-level schemes and according to an induced semantic flip rate for sentence-level schemes. These contraction laws reveal an irreducible trilemma among robustness, stealth, and reliable verification. Guided by these limits, we propose a hybrid watermarking strategy that selects among distribution-preserving, semantic-level, and token-level methods based on anticipated editing regimes. Experiments on Llama-2-7B and Mistral-7B under paraphrasing attacks corroborate theoretical predictions and confirm that the hybrid strategy is empirically near-Pareto across evaluated edit regimes.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

The recent emergence of LLM watermarking has brought a persistent phenomenon into sharp focus: a detectability and robustness trade-off is consistently observed across watermarking families (Fig. 1). Early token-level probability-modifying watermarks, including biased "green-list" schemes in the KGW lineage (Kirchenbauer et al., 2023) and robust unigram-style variants (Zhao et al., 2024), achieve strong resilience to downstream editing but do so by introducing statistical drift that can be detected even by *keyless* observers. This pressure toward greater stealth has motivated semantic and sentence-level schemes such as SemStamp, PMark, and SimMark (Hou et al., 2024; Huo et al., 2025; Dabiriaghdam & Wang, 2025), which encode evidence at coarser granularity to better survive paraphrasing while substantially reducing exposure to token-level detection tests.

The observed trade-off has further driven the development of schemes pursuing joint watermark and detector design to approach optimal operating points: distribution-adaptive watermarking algorithm (DAWA) explicitly couples the generator-side rule and detection statistic (He et al., 2025), while HeavyWater and SimplexWater target min-max optimality in low-entropy next-token regimes (Tsur et al., 2025). Beyond inference-time sampling, structural or training-time watermarks aim to be robust and stealthy while maintaining utility by embedding signals into model parameters or training dynamics (Gu et al., 2024; Block et al., 2025). At the extreme end, cryptographic distribution-preserving watermarks (Christ et al., 2024) offer perfect undetectability to keyless observers by construction but are brittle under even mild output editing, since verification depends on preserving fine-grained token alignment.

A key limitation of prior work lies in how the detectability and robustness trade-off is quantified. Existing approaches rely primarily on detector-specific empirical statistics, notably black-box tests and calibrated scores such as z-scores and p-scores, measured before and after paraphrasing or other editing attacks (Gloaguen et al., 2025; Liu et al., 2025; Li et al., 2025). While these tests capture observable properties of watermarked text, *they do not provide a unified theoretical standpoint that can (i) compare heterogeneous*

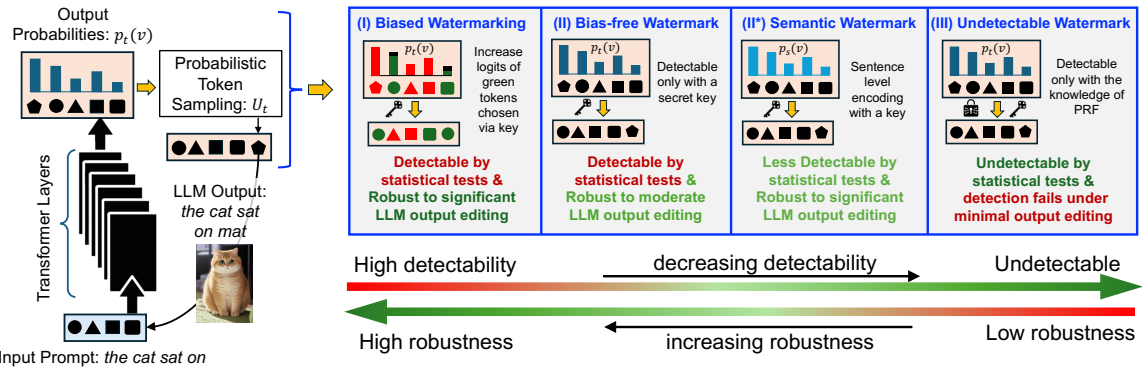


Figure 1. Watermarking schemes in modern LLMs exhibit a trade-off between detectability via statistical tests and robustness against LLM output editing.

watermark families under a common currency and (ii) tie a scheme's robustness and keyless detectability directly to weaknesses inherent in its construction principles.

In this work, we develop an information-theoretic framework for inference-time watermarking that assigns the trade-off to four watermark families: biased token-level, unbiased token-level, semantic-level, and distribution-preserving schemes. The limits are information-theoretic; for token-level schemes, likelihood-ratio detectors are efficient given oracle or surrogate access to model probabilities and the watermark rule, whereas for sample-only keyless outsiders, these bounds need not be tight. We formalize detector access models in Section 3. This provides a common language for analyzing existing proposals and guiding a hybrid strategy empirically near the Pareto boundary. We focus on the accumulated KL divergence, the *usable information budget*, which upper-bounds stealth via Pinsker's inequality and lower-bounds verification power via Neyman-Pearson bounds. This budget contracts under editing, assuming an idealized noise model, yielding a trilemma for fixed-vocabulary LLMs. We empirically validate the predicted contraction trends under our evaluated paraphrasing and back-translation attacks, while noting that structured edits can deviate from the idealized model.

Our framework proceeds in two steps. First, we measure detectability using total variation distance, establishing a hierarchy across token-level and semantic-level schemes (Theorem 3.2). Second, we show how the KL budget shrinks under editing (Theorem 3.4). For token-level schemes, the information-theoretic limit can be achieved efficiently when the detector has access to the model probabilities and the watermark rule, and, for keyed schemes, also to the key. For semantic schemes, optimal detection requires acceptance probabilities a_t. Building on these results, we derive a *minimal-information hybrid selection rule* that chooses among watermark families based on anticipated edits and stealth requirements (Theorem 4.2). Experiments on Llama and Mistral confirm that the hybrid strategy is near Pareto in post-edit verification across evaluated regimes.

Our principal contributions are as follows:

1. **Detectability and post-edit verification characterization:** We show that a single KL quantity governs both stealth (upper-bounded via Pinsker) and robustness after edits (lower-bounded via Neyman-Pearson). This budget contracts with edit rate, yielding explicit feasibility regions (Theorem 3.4).
2. **Minimal-information hybrid selection rule:** We derive a scheme selecting among watermarking families based on anticipated edits, choosing the method requiring the least KL budget to meet target verification power subject to a stealth cap (Theorem 4.2).
3. **Experimental validation:** We confirm theoretical predictions through paraphrasing attacks on Llama and Mistral, demonstrating near-Pareto post-edit verification at 15-30% edit rates while maintaining a Pinsker TV bound below 0.1.

The remainder of this paper is organized as follows. Section 2 reviews existing watermarking approaches. Section 3 develops our information-theoretic framework. Section 4 derives the minimal-information hybrid selection rule. Section 5 experimentally validates our theoretical predictions. Finally, Section 6 concludes the paper.

2. Related Works on LLM Watermarking and Research Gap

Inference-time watermarking schemes for LLMs can be organized along two axes: (i) *granularity* ranging from token-level to sentence-level, and (ii) whether the watermark *modifies* the output distribution for a fixed key or preserves it exactly. We adopt this organization to position prior work and motivate the unified analysis in Section 3, deferring detailed technical treatment to Appendix A. Existing token-level watermarking schemes modify the generation process through three distinct approaches:

1. **Biased sampling** (Kirchenbauer et al., 2023; Zhao et al., 2024) designates certain tokens as "green" at each gen-

eration step and applies an exponential tilt to sampling probabilities. These schemes achieve strong empirical robustness (Kirchenbauer et al., 2024) but remain readily detectable through statistical tests (Gloaguen et al., 2025; Liu et al., 2025) and truncated goodness-of-fit methods (Li 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, satisfying $\mathbb{E}_E[R_E(p_t)] = p_t$. Despite this first-order unbiasedness, all such schemes remain detectable through variance analysis (Gloaguen et al., 2025). Recent variants such as HeavyWater and SimplexWater (Tsur et al., 2025) formulate watermark design as minimax optimization, achieving improved detection in low-entropy regimes.
3. **Distribution-preserving sampling** (Christ et al., 2024; Zamir, 2024) maintains exact token probabilities using pseudorandom functions. These schemes achieve provable undetectability but fail catastrophically when LLM outputs are perturbed.

A recent extension by Golowich & Moitra (2024) introduces substring robustness while preserving undetectability; however, this approach requires vocabulary sizes polynomial in the security parameter, with degree $\Theta\big(\frac{1}{\alpha}\log\frac{1}{\alpha}\big)$ in an entropy-rate parameter α. For realistic entropy levels and constant-fraction edit robustness, this requirement exceeds the vocabulary sizes of practical LLMs (see Appendix A.6).

Beyond token-level methods, researchers have explored coarser granularities to improve robustness against meaning-preserving attacks. **Semantic and sentence-level watermarking** methods such as SemStamp (Hou et al., 2024), PMark (Huo et al., 2025), SIR (Liu et al., 2024), and SimMark (Dabiriaghdam & Wang, 2025) embed watermark evidence through keyed selection rules operating in embedding or proxy spaces. By operating at the sentence level rather than the token level, these approaches reduce sensitivity to surface-level rewrites and synonym substitutions. All of these approaches fall within the probability-modifying category: for any fixed key, the selection rule induces a conditional distribution that differs from the base model. Consequently, while they improve robustness, they do not eliminate the detectability-robustness trade-off (Theorem 3.4).

Operating at an even coarser granularity, **generator-side selection** approaches such as WaterMax (Giboulot & Furon, 2024) generate multiple candidate continuations and select the output that maximizes a keyed watermark criterion. These methods shift the cost from logit manipulation to increased computation, yet the selection operation still induces a distributional shift at the whole-text level, placing them within the probability-modifying family. WaterMax operates as a meta-method over existing watermarking schemes, inheriting their underlying detectability characteristics. While our taxonomy concerns inference-time schemes, **training-time methods** such as GaussMark (Block et al., 2025) embed watermarks via weight perturbations; these remain detectable insofar as they induce shifts in the output distribution. We include GaussMark in our evaluation (Section 5) as a representative structural baseline.

Table 1. Notation guide for the main quantities used in Section 3.

Symbol / Term	Meaning
T	Number of tokens in the generated text.
T_s	Number of sentences in the generated text.
$p_t(\cdot)$	Baseline conditional distribution at step t.
$q_t(\cdot)$	Watermarked conditional distribution at step t.
P^s	Sequence distribution induced by the baseline sampler.
Q	Sequence distribution induced by the watermarked sampler.
δ	Bias strength in biased token-level sampling.
g_t	Baseline green-set mass at step t.
$\bar{\sigma}^2$	Variance term for bias-free token-level watermarking.
Z_t	Semantic evidence bit at sentence step t.
ρ	Semantic bias parameter controlling watermark strength.
D_0	Per-unit information budget before edits.
$\mathrm{TV}(P^s, Q)$	Total variation between the baseline and watermarked distributions.
$\mathrm{KL}(Q\|P^s)$	KL divergence from the watermarked distribution to the baseline distribution.
$\mathrm{Detect}_{\mathrm{IT}}$	Information-theoretic detectability.
$\mathrm{Detect}_{\mathrm{comp}}$	Computational detectability over efficient detectors.
keyless observer	Detector that only observes generated text, without access to the key, model probabilities, or watermark rule.
D_λ	A randomized detector operating at security / problem size λ.
$\mathrm{Adv}_{D_\lambda}(\lambda)$	Distinguishing advantage of detector D_λ, i.e., how well it separates watermarked from unwatermarked text.
$\mathrm{ED}(y, \tilde{y})$	Edit distance between original text y and edited text $\tilde{y}$.
ε	Token edit rate after post-generation editing.
ε_s	Probability that editing flips the semantic evidence bit.
$C_{\mathrm{tok}}(\varepsilon)$	Usable post-edit information budget for token-level watermarking.
$C_{\mathrm{sem}}(\varepsilon)$	Usable post-edit information budget for semantic watermarking.
α	False-alarm level of the detector.
β	Target miss probability of the detector.
$D_{\mathrm{req}}^{\mathrm{tok}}$	Required information budget for token-level schemes under the given edit regime.
$D_{\mathrm{req}}^{\mathrm{sem}}$	Required information budget for semantic schemes under the given edit regime.
ℓ	Average number of tokens per sentence.

Research gap. Despite rapid progress, existing analyses remain fragmented across scheme families and typically do not provide: (i) a *cross-family detectability hierarchy* under a unified metric enabling comparison of token-level, sentence-level, and distribution-preserving approaches; (ii) a principled characterization of how post-generation *editing reduces usable detection information*, beyond heuristic robustness measures; or (iii) an explicit *design rule* for selecting among scheme families under specified output editing regimes and stealth constraints. Our framework addresses these gaps by relating detectability to distributional distance through a KL budget and robustness to the contraction of detection-relevant information under editing, thereby enabling principled hybrid selection.

3. Robustness vs. Detectability Trade-off

The detectability and robustness of watermarked text depend fundamentally on how outputs are sampled during generation. When a language model generates text, it computes conditional probability distributions and applies a

sampling rule that converts these probabilities into realized outputs. The resulting distribution over complete texts depends on the model probabilities, the sampling mechanism, and the secret key when present. This section develops an information-theoretic framework quantifying detectability and robustness for both token-level and semantic watermarking schemes, establishing the fundamental trade-offs constraining all probability-modifying watermarks.

We define our central information quantity, the *usable KL budget*, as the accumulated KL divergence between the watermarked distribution and the baseline, which governs the separability of hypothesis testing. Unless stated otherwise, we measure KL divergence in *bits*.

Randomness enters generation at each step t.[1] In the token-level setting, the model provides a conditional distribution $p_t(\cdot) = p_\theta(\cdot \mid x, \mathbf{y}_{<t})$ over vocabulary Σ, where x denotes the prompt and $\mathbf{y}_{<t}$ denotes the previously generated tokens. A sampling rule s uses $U_t \sim \mathrm{Uniform}[0,1]$ (and possibly secret keys) to produce the next token y_t. In the semantic setting (PMark, SemStamp), the sampler generates entire sentences $s_t \in \Sigma^\star$ at each step, with watermarking applied via keyed semantic selection rules defined through encoders, locality-sensitive hashes, or proxy models. In both settings, watermarked sampling can modify either the induced conditional distribution (creating $q_t \neq p_t$ for a fixed key) or the randomness source, or both.

Definition 3.1 (Detectability)**.** Let s denote a baseline sampling rule inducing distribution P^s over texts Ω, and let $\tilde{s}$ be a keyed watermarked rule inducing $Q^{\tilde{s}}$. The **information-theoretic detectability** is $\mathrm{Detect}_{\mathrm{IT}}(\tilde{s}) := \mathrm{TV}(P^s, Q^{\tilde{s}})$, with Pinsker's inequality providing the upper bound $\mathrm{Detect}_{\mathrm{IT}}(\tilde{s}) \leq \sqrt{\frac{1}{2}\mathrm{KL}(Q^{\tilde{s}} \| P^s)}$. The **computational detectability** is $\mathrm{Detect}_{\mathrm{comp}}(\tilde{s}_\lambda) := \sup_{D_\lambda \in \mathsf{PPT}} \mathrm{Adv}_{D_\lambda}(\lambda)$, where the supremum is over probabilistic polynomial-time detectors. In experiments, we report $\sqrt{\frac{1}{2}\mathrm{KL}(Q \| P)}$ as a Pinsker upper bound on TV, not a direct TV estimate.

Detector Access Models

An **oracle** detector has access to true conditionals $p_t(\cdot \mid x, y_{<t})$, the watermark rule, and the key if acting as verifier. A **surrogate** detector uses an independently trained model $\hat{p}_t$. A **sample-only** detector observes text samples without probability.

Information-theoretic detectability captures the best distinguishing advantage achievable regardless of computational resources, while computational detectability restricts attention to efficient algorithms. By construction, $\mathrm{Detect}_{\mathrm{comp}} \leq \mathrm{Detect}_{\mathrm{IT}}$ (Lemma C.2). Equality holds for oracle or surrogate detectors that can evaluate the likelihood ratio; otherwise only the upper bound is asserted. Under sample-only access, Pinsker-based bounds may not be tight.

[1] All notation is summarized in Appendix B.

3.1. Detectability Characterization

We establish quantitative bounds on information-theoretic detectability for each sampling family with fixed key constructions. Key-averaged variants appear in Appendix C.7.

Theorem 3.2 (Information-theoretic detectability)**.** *Fix a prompt x and token length T, and let P^s denote the baseline distribution induced by standard stochastic sampling. The total variation distance $\mathrm{TV}(P^s, Q)$ satisfies the bounds given in Table 2.*

In the table, Q^{greedy} denotes the point mass on the deterministic greedy sequence $\mathbf{y}^\star$, Q^{bias_δ} denotes the tilted distribution with parameter δ applied to a keyed green set G_t where $g_t = p_t(G_t)$ is the baseline green-set mass, Q_E^{bf} is obtained from an unbiased reweighting operator R_E satisfying $\mathbb{E}_E[R_E(p_t)] = p_t$, Q_k^{sem} is induced by keyed sentence-level semantic selection with acceptance mass a_t at sentence step t, and Q^{prf} preserves $q_t \equiv p_t$ at every step. In typical regimes, the relation $T \approx \ell T_s$ holds, where ℓ denotes the average number of tokens per sentence. The proof is provided in Appendix C.

Informal Summary of Theorem 3.2

Theorem 3.2 says that if a watermark changes how the model samples text, then it also leaves a detectable statistical trace. That trace grows with the length of the generated text, so longer outputs are easier to distinguish from ordinary model samples. The effect is strongest for greedy and biased schemes, weaker for semantic schemes, and absent only for distribution-preserving schemes.

Interpretation of Theorem 3.2

- **Accumulation over length.** For probability-modifying schemes, $\mathrm{TV}(P^s, Q)$ grows with generation length and watermark parameters, reflecting accumulated statistical evidence.
- **Detectability hierarchy.** Greedy exhibits $O(1)$, biased grows as $O(|\delta|\sqrt{T})$, bias-free as $O(\sqrt{T})$, and distribution-preserving achieves $\mathrm{TV} = 0$.
- **Semantic schemes.** Sentence-level semantic selection accumulates as $O(\sqrt{T_s})$, corresponding to $O\left(\sqrt{T/\ell}\right)$ under typical sentence length ℓ.

For any probability-modifying watermark with fixed key k, we have $Q_k \neq P^s$ and thus $\mathrm{TV}(P^s, Q_k) > 0$, so an information-theoretic distinguisher always exists. In the non-cryptographic setting, the induced statistical drift admits efficiently computable test statistics, consistent with recent statistical detectors achieving practical success against biased and bias-free watermarks (Gloaguen et al., 2025). Theorem 3.2 explains why these detectors become increasingly powerful as length grows, establishing the foundation for the robustness-detectability trade-off analyzed next.

Table 2. Information-theoretic detectability bounds for fixed-key sampling methods.

Sampling family	TV bound	Inline meaning
Greedy	$\mathrm{TV}(P^s, Q^{\mathrm{greedy}}) = 1 - P^s(\mathbf{y}^\star) = O(1)$	Q^{greedy}: point mass on $\mathbf{y}^\star$
Biased (δ-tilt)	$\mathrm{TV}(P^s, Q^{\mathrm{bias}_\delta}) \le \lvert\delta\rvert\sqrt{\frac{1}{4}\sum_{t=1}^{T} g_t(1-g_t)} = O(\lvert\delta\rvert\sqrt{T})$	δ: bias strength; g_t: green-set mass
Bias-free (key E)	$\mathrm{TV}(P^s, Q_E^{\mathrm{bf}}) \le \sqrt{\frac{1}{4}\sum_{t,v}\frac{\mathrm{Var}_E[R_E(p_t)(v)]}{p_t(v)}} = O(\sqrt{T})$	E: key; R_E: reweighting operator
Semantic (key k)	$\mathrm{TV}(P^s, Q_k^{\mathrm{sem}}) \le \sqrt{\frac{1}{2}\sum_{t=1}^{T_s}\mathbb{E}\left[\log\frac{1}{a_t}\right]} = O\left(\sqrt{T/\ell}\right)$	a_t: acceptance mass; $T \approx \ell T_s$
Distribution-preserving	$\mathrm{TV}(P^s, Q^{\mathrm{prf}}) = 0$	$q_t \equiv p_t$

P^s denotes the baseline text distribution induced by standard stochastic sampling, and Q denotes the corresponding watermarked text distribution.
T is the number of token-generation steps. In the semantic setting, T_s is the number of sentence-generation steps and ℓ is the average number of tokens per sentence, so typically $T \approx \ell T_s$.
In the greedy row, $\mathbf{y}^\star$ is the deterministic greedy output sequence and Q^{greedy} is the point mass on that sequence.
In the biased row, δ is the logit-tilt strength, G_t is the keyed green set at step t, and $g_t := p_t(G_t)$ is the baseline probability mass assigned to that set.
In the bias-free row, E denotes the watermark key or seed, and R_E is the reweighting operator satisfying $\mathbb{E}_E[R_E(p_t)] = p_t$.
In the semantic row, a_t is the acceptance probability (acceptance mass) at sentence step t under the keyed semantic selection rule.
In the distribution-preserving row, $q_t \equiv p_t$ means that the conditional token distribution is preserved exactly at every generation step.
All bounds are stated for fixed-key constructions.

3.2. Robustness Analysis Under Text Perturbations

The fundamental tension in watermarking lies in balancing *stealth* (low detectability to keyless outsiders) with *robustness* (reliable verification by key holders after edits). We quantify stealth via KL divergence and total variation, and robustness via detection power at miss probability β.

Definition 3.3 (Robustness)**.** Fix length T, edit tolerance $\varepsilon \in [0,1]$, and false-alarm level $\alpha \in (0,1)$. The family $\{\tilde{s}_\lambda\}$ is $(\varepsilon, \alpha, \beta)$-information-theoretically robust if there exists a level-α detector achieving power at least $1-\beta$ on edited watermarked text satisfying $\mathrm{ED}(y, \tilde{y}) \le \varepsilon T$. Computational robustness is defined analogously with the supremum restricted to PPT detectors.

3.2.1. Edit Channel Model

We model edits via a mixture (substitution) channel: at each position, the original token is retained with probability $1-\varepsilon$ and replaced with probability ε by a draw from a replacement distribution. Our derivations use the i.i.d. simplification to obtain closed-form contractions, but the key $(1-\varepsilon)^2$ attenuation for token-level KL budgets arises for any mixture channel that linearly attenuates the perturbation (see Appendix D). This channel serves as a tractable first-order model whose parameter ε can be calibrated to match empirical edit rates observed under specific attack scenarios. We empirically validate that the predictions remain accurate under correlated paraphrasing edits in Section 5.

3.2.2. Per-unit KL Budgets and Usable Information After Edits

We establish the per-unit information budgets and effective post-edit usable KL budgets for both watermarking families. Our quadratic KL approximations assume that per-step perturbations are small. For biased sampling, the tilt parameter δ controlling the probability shift toward the green set satisfies $|\delta| \ll 1$, and the green-set mass $g_t := p_t(G_t)$ remains bounded away from $\{0,1\}$. For bias-free reweighting, the perturbation $\epsilon_{t,E}(v)$ to the probability of token v at step t under key E satisfies $|\epsilon_{t,E}(v)| \le \eta\, p_t(v)$ for some small $\eta \ll 1$, where $p_t(v)$ denotes the baseline probability. For semantic watermarking, the bias parameter ρ controlling the deviation of the sentence-level indicator from uniform satisfies $|\rho| \ll 1$.

Token-level information budgets. The per-token information is $D_0^{(\mathrm{biased})} \approx \delta^2 g_t(1-g_t)/(2\ln 2)$ for biased sampling and $D_0^{(\mathrm{bias\text{-}free})} \approx \hat{\sigma}^2/(2\ln 2)$ for bias-free sampling, where $\hat{\sigma}^2 = \sum_v p_t(v)\mathrm{Var}_E[R_E(v)]$ is the reweighting variance. The post-edit usable KL budget contracts as $C_{\mathrm{tok}}(\varepsilon) \approx T(1-\varepsilon)^2 D_0^{(\mathrm{tok})}$.

Semantic information budgets. Semantic watermarks reduce each sentence S_t to a keyed binary indicator $Z_t := g_k(F(S_t)) \in \{0,1\}$, with $\mathbb{E}[Z_t \mid H_0] = \frac{1}{2}$ under the null and $\mathbb{E}[Z_t \mid H_1] = \frac{1}{2} + \rho$ under watermarking. The per-sentence information is $D_0^{(\mathrm{sem})} \approx 2\rho^2/\ln 2$. Semantic evidence degrades when token edits flip the sentence-level indicator, captured by the induced semantic flip rate $\varepsilon_s(\varepsilon) := \Pr[g_k(F(\widetilde{S}_t)) \ne g_k(F(S_t))]$, yielding the post-edit budget $C_{\mathrm{sem}}(\varepsilon) \approx T_s(1-2\varepsilon_s(\varepsilon))^2 D_0^{(\mathrm{sem})}$.

Comparing the two channels. Token-level schemes experience direct attenuation proportional to $(1-\varepsilon)^2$, whereas semantic schemes experience attenuation through $\varepsilon_s(\varepsilon)$. This decoupling enables semantic schemes to remain robust under paraphrasing attacks that modify many tokens while preserving meaning.

3.2.3. Unified Robustness-Detectability Trade-off

Theorem 3.4 (Robustness-Detectability Trade-off)**.** *Fix T tokens, T_s sentences, edit rate ε, and false-alarm level $\alpha \in (0,1)$. For a level-α Neyman-Pearson test, achieving target miss probability β requires $C(\varepsilon) \gtrsim \log_2(1/\beta)$. This yields the maximal tolerable edit rate for token-level schemes: $\varepsilon_\beta^{\mathrm{tok}} = 1 - \sqrt{\log_2(1/\beta)/(TD_0^{(\mathrm{tok})})}$, and the constraint on semantic flip rate: $\varepsilon_s(\varepsilon) \le \frac{1}{2}(1 -$*

$\sqrt{\log_2(1/\beta)/(T_s D_0^{(\text{sem})})}$).

The proof (Appendix D) proceeds via per-unit KL expansions, quadratic contraction under edits, and the chain rule. In practice, ε can be measured empirically, and $\varepsilon_s(\varepsilon)$ can be estimated by evaluating embedding stability under perturbation, as we demonstrate in Section 5.

Informal Summary of Theorem 3.4

Theorem 3.4 says that editing makes watermark verification harder by removing part of the watermark signal. For token-level methods, this loss depends directly on the number of edited tokens. For semantic methods, it depends on how often editing changes the sentence-level watermark evidence. If too much signal is lost, reliable verification becomes impossible. This is why token-level methods perform best under light editing, whereas semantic methods are better when wording changes substantially while the sentence-level signal remains intact.

Interpretation of Theorem 3.4

- **Quadratic contraction.** Robustness degrades as $(1-\varepsilon)^2$ for token-level and $(1-2\varepsilon_s)^2$ for semantic schemes.
- **Semantic resilience.** When attacks exhibit high ε but low $\varepsilon_s(\varepsilon)$, semantic schemes retain substantially more signal.
- **Stealth-robustness tension.** Stealth requirements force small D_0, reducing the tolerable corruption level.

3.3. Implications for Watermark Design

Theorem 3.4 provides a design rule: robustness improves by increasing redundancy (T or T_s) and per-unit budget (D_0), and degrades quadratically with effective edit rate. Token-level schemes are advantageous when edits are uniformly distributed and ε is low. Semantic schemes become preferable when attacks exhibit high surface-level modification but preserve semantic content. The ratio $\varepsilon_s(\varepsilon)/\varepsilon$ serves as a diagnostic for selecting between families.

Corollary 3.5 (Impossibility Region)**.** *Fix length T, watermark strength $D_0^{(\text{tok})}$, and target power $1-\beta$. For edit rates $\varepsilon > \varepsilon_\beta^{\text{tok}}$, reliable detection is unattainable for any probability-modifying watermark with parameters $(T, D_0^{(\text{tok})})$, even for an information-theoretic detector.*

Corollary 3.5 formalizes the design dilemma: one cannot simultaneously achieve large edit tolerance and guaranteed verification. For instance, with $T = 500$, $D_0^{(\text{tok})} = 0.02$ bits/token, and $\beta = 0.01$, the maximal tolerable edit rate is $\varepsilon_\beta^{\text{tok}} \approx 0.18$.

This trade-off raises a design question: *given an anticipated edit regime, how should one select watermark parameters?* Section 4 formalizes this through a composite loss incorporating reliability, detectability, and parameter efficiency. Minimizing this loss yields a closed-form hybrid strategy that allocates across distribution-preserving, semantic, and token-level methods based on the operating regime.

4. Optimal Watermark Selection Under Output Editing

Theorems 3.2 and 3.4 jointly imply a design principle: to achieve target verification power under an anticipated edit regime while remaining stealthy, one should operate at the smallest information level that meets the robustness requirement. This section develops a practical family-selection rule across distribution-preserving, token-level, and semantic watermarking schemes. The full optimization technique and proofs are deferred to Appendix E.

4.1. A Design Rule for Watermark Selection

Fix detector level α and miss probability β, so that the test has power $1-\beta$. By the Neyman-Pearson sufficiency logic underlying Theorem 3.4, reliable detection requires that the available information exceed $\log_2(1/\beta)$. Using the channel capacities from Section 3.2.2, we derive the minimum per-unit information required to achieve target power:

$$D_{\text{req}}^{\text{tok}}(\varepsilon, T, \beta) := \frac{\log_2(1/\beta)}{T(1-\varepsilon)^2}, \tag{1}$$

$$D_{\text{req}}^{\text{sem}}(\varepsilon, T_s, \beta) := \frac{\log_2(1/\beta)}{T_s(1-2\varepsilon_s(\varepsilon))^2}. \tag{2}$$

Stealth requirements constrain these budgets from above. If an outsider pools M tokens and requires a total variation bound $\text{TV} \le \tau$, then Pinsker's inequality yields the KL-based cap

$$D_0^{(\text{tok})} \le D_{\text{stealth}}^{\text{tok}}(M, \tau) := \frac{2\tau^2}{M \ln 2}. \tag{3}$$

Analogously, pooling M_s sentences with budget τ_s yields

$$D_0^{(\text{sem})} \le D_{\text{stealth}}^{\text{sem}}(M_s, \tau_s) := \frac{2\tau_s^2}{M_s \ln 2}. \tag{4}$$

These constraints define feasibility regions for each watermarking family. A scheme is feasible if its required information level does not exceed the stealth cap, that is, $D_{\text{req}} \le D_{\text{stealth}}$ for the respective parameters.

Definition 4.1 (Hybrid Watermark Selection Rule)**.** Fix (ε, T), $(\varepsilon_s(\varepsilon), T_s)$, target miss probability β, and stealth parameters (M, τ), (M_s, τ_s). The hybrid selection rule $\mathcal{H}$ proceeds as follows:

(i) If a distribution-preserving watermark with K marked positions achieves $\Pr[X < t] \le \beta$ under $X \sim \text{Binomial}(K, 1-\varepsilon)$, select it with $D_0 = 0$. The case $\varepsilon = 0$ falls within this region.

(ii) Otherwise, compute $D_{\text{req}}^{\text{tok}}$ and $D_{\text{req}}^{\text{sem}}$ from (1)–(2) and compare with the corresponding D_{stealth}.

(iii) If both families are feasible, select semantic watermarking if

$$D_{\text{req}}^{\text{sem}}(\varepsilon, T_s, \beta) < D_{\text{req}}^{\text{tok}}(\varepsilon, T, \beta), \tag{5}$$

and select token-level watermarking otherwise.

(iv) If only one family is feasible, select that family.

(v) Set D_0 to the minimum feasible value D_{req} for the selected family.

The selection rule in (5) admits a simple interpretation: semantic watermarking is preferred when

$$\frac{(1-2\varepsilon_s(\varepsilon))^2}{(1-\varepsilon)^2} > \frac{T_s}{T}, \tag{6}$$

that is, when the semantic channel retains a larger fraction of its original capacity than the token-level channel, adjusted for differences in unit counts. This condition is satisfied when attacks induce high token edit rates but rarely flip semantic indicators, as occurs with synonym substitution and meaning-preserving paraphrasing.

Theorem 4.2 (Optimality of Hybrid Selection)**.** *Among schemes in the considered family class satisfying the Pinsker-based stealth constraints (3)–(4) and achieving power $1-\beta$ at level α under edit rate ε, the hybrid rule $\mathcal{H}$ minimizes the KL-derived TV bound to keyless adversaries. The selected family achieves the smallest required information budget, and setting $D_0 = D_{\text{req}}$ attains minimum stealth cost while meeting the robustness target. If neither family is feasible, no scheme in this class achieves the target power.*

The proof appears in Appendix E. Unlike token-level schemes whose robustness depends directly on ε, semantic schemes depend on $\varepsilon_s(\varepsilon)$, which varies with both the watermark design and the attack strategy. In practice, $\varepsilon_s(\varepsilon)$ can be estimated offline by applying the expected editing pipeline to representative outputs and measuring the empirical flip frequency. To guard against estimation error, a conservative selector can use upper-confidence bounds $(\varepsilon^{\text{U}}, \varepsilon_s^{\text{U}})$ when computing D_{req}, yielding a high-probability guarantee that the selected family meets the target power.

Informal Summary of Theorem 4.2

Theorem 4.2 states that the best watermark family is the feasible one that requires the smallest information budget to meet the robustness target. If a distribution-preserving scheme is sufficiently robust, it is optimal because it achieves the target without statistical drift. Otherwise, the selector compares the required budgets for token-level and semantic watermarking and selects the smaller feasible option. If neither is feasible, then the desired operating point cannot be achieved within this family class.

5. Experimental Evaluation

This section empirically validates our information-theoretic framework using four families of *inference-time* watermarking schemes, evaluating both detectability and robustness against paraphrasing attacks. These families comprise *biased* token-level sampling, *bias-free* token-level sampling, *semantic* (sentence-level) rejection or selection methods, and *distribution-preserving* sampling. In addition to these core families, we include several recent watermarking methods that extend or complement our hierarchy: (i) **GaussMark** (Block et al., 2025), a *training-time* watermark embedded in model weights; (ii) **HeavyWater** and **SimplexWater** (Tsur et al., 2025), token-level designs that achieve bias-free behavior in expectation for low-entropy next-token distributions; and (iii) **SimMark** (Dabiriaghdam & Wang, 2025), a sentence-level similarity watermark. Beyond robustness and keyless detectability, practical deployments require preserving output utility and controlling runtime overhead. We therefore evaluate watermarking schemes along three axes: (i) *robustness* of keyed verification after editing (AUROC, TPR at 1% FPR), (ii) *keyless detectability* by black-box statistical detectors, and (iii) *utility and cost* (MAUVE, BERTScore, LLM-as-judge, and compute overhead). Unless stated otherwise, all reported numbers are means over prompts with 95% bootstrap confidence intervals.

All the relevant codes and a detailed user manual for replicating the experiments in this work are available at `https://anonymous.4open.science/r/Catch-22-Pareto-Frontier-Watermark-in-LLMs-040B` and will be released upon acceptance.

Experimental Setup

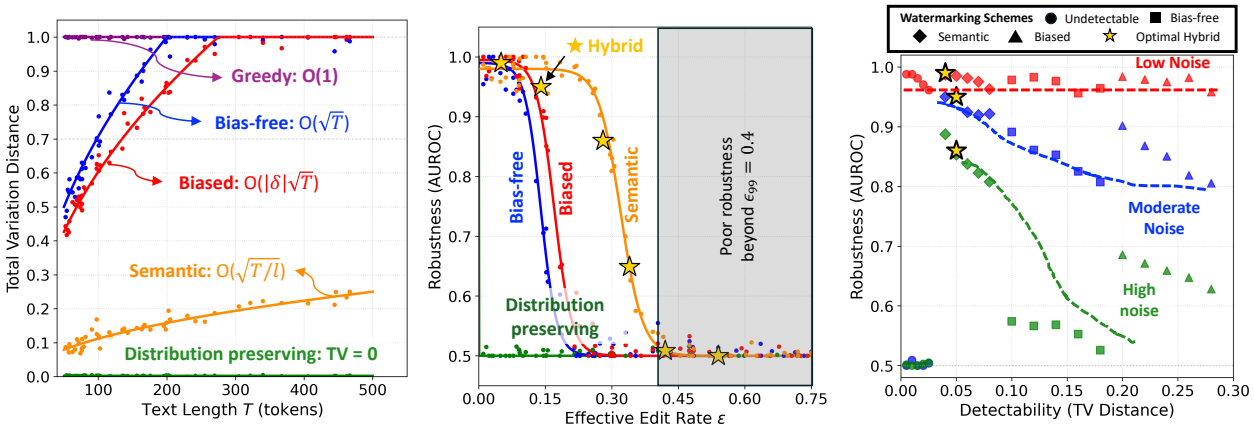
Dataset and Models. We use 500 prompts from the LFQA dataset (Krishna et al., 2023), generating 100 to 1000 tokens per response with Llama-2-7B (Touvron et al., 2023) on a single NVIDIA H100 GPU. Mistral-7B (Jiang et al., 2023) results appear in Appendix G.

Watermarking Schemes. We evaluate methods from each inference-time family: *biased* (KGW, Unigram (Kirchenbauer et al., 2023; Zhao et al., 2024)), *bias-free* (DiPMark, HCW (Wu et al., 2024; Hu et al., 2024)), *semantic* (SemStamp, PMark (Hou et al., 2024; Huo et al., 2025)), and *distribution-preserving* (CGW (Christ et al., 2024)). We additionally include HeavyWater and SimplexWater (Tsur et al., 2025), SimMark (Dabiriaghdam & Wang, 2025), and GaussMark (Block et al., 2025). Our **Hybrid** scheme (Theorem 4.2) dynamically selects among families based on edit characteristics.

Attacks. Oblivious paraphrasers: DIPPER (Krishna et al., 2023) ($\hat{\varepsilon} \approx 0.25$), OPT-2.7B (Zhang et al., 2022) ($\hat{\varepsilon} \approx 0.15$), synonym substitution ($\hat{\varepsilon} \approx 0.15$), watermark-removal prompting ($\hat{\varepsilon} \approx 0.15$), and back-translation (Liu et al., 2025) ($\hat{\varepsilon} \approx 0.42$). Adaptive attacks are referred in Appendix G.

Metrics. Keyed robustness: AUROC, TPR at 1% FPR. Keyless detectability: z-score detector (Liu et al., 2025); Tr-GoF detector (Li et al., 2025) discussed in Appendix G. Utility: MAUVE, BERTScore, LLM-as-judge (Appendix G).

Table 3 presents comprehensive results on Llama-2-7B, reporting keyed robustness and keyless detectability across six conditions: no attack, two paraphrasers at moderate edit



(a) *(b)* *(c)*

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

rates (DIPPER and OPT-2.7B), two lexical attacks (synonym substitution and watermark-removal paraphrasing), and back-translation at a high edit rate. Results for Mistral-7B, adaptive attack evaluation, utility metrics, and additional analyses appear in Appendix G.

5.1. Key Experimental Findings

Robustness and detectability tradeoffs sharpen with edit rate. Under reference conditions without paraphrasing, all schemes achieve near-perfect robustness (AUROC ≥ 0.98), yet detectability varies substantially across families. Biased schemes exhibit high z-scores (11.2 to 30.1), making them readily identifiable by (Liu et al., 2025). CGW is not flagged by our keyless test (scores near the unwatermarked baseline), consistent with its distribution-preserving design. As edit rates increase, all families move away from the empirical Pareto frontier: performance degrades consistently with the predicted contraction (Theorem 3.4), manifesting as drops in AUROC and TPR. This pattern is most pronounced under back-translation ($\hat{\varepsilon} \approx 0.42$), where even bias-free schemes collapse to near-random performance (AUROC of 0.55 to 0.60).

Semantic watermarks achieve higher verification metrics in moderate-edit regimes but degrade under extreme paraphrasing. Semantic schemes demonstrate superior robustness when token-level edits preserve underlying meaning. Under the DIPPER attack ($\hat{\varepsilon} = 0.25$), PMark achieves AUROC = 0.94 versus 0.91 for HCW; under OPT-2.7B ($\hat{\varepsilon} = 0.15$), this gap widens to 0.95 versus 0.92. These results are consistent with Theorem 3.4: semantic schemes excel when the token edit rate ε is high but the induced semantic flip rate $\varepsilon_s(\varepsilon)$ remains low. However, under back-translation ($\hat{\varepsilon} \approx 0.42$), even semantic watermarks degrade substantially, with AUROC falling to 0.80 to 0.85. This supports the paper's central message: maintaining verification at high edit rates generally requires stronger embedding, which can increase detectability.

The hybrid scheme matches or exceeds the best single-family method across conditions. The hybrid scheme from Theorem 4.2 matches or exceeds the best evaluated baseline in most conditions while maintaining low detectability. Under DIPPER, the hybrid matches PMark (AUROC = 0.94) by adaptively selecting semantic watermarking when paraphrasing preserves meaning but alters surface tokens. Under back-translation, the hybrid achieves AUROC = 0.86, selecting the best feasible point when no evaluated method simultaneously achieves high verification and low detectability. Sensitivity analysis shows that the hybrid selector depends on accurate estimates of ε and $\varepsilon_s(\varepsilon)$; a conservative selector using upper confidence bounds reduces worst-case regret with negligible loss in average performance (Appendix G).

Recent baselines reinforce rather than overturn the observed tradeoff. GaussMark (Block et al., 2025) embeds a structural watermark at training time; although it lies outside our inference-time hierarchy, it still shifts the output distribution and is detectable, achieving z-score of 12.4 in the reference condition. HeavyWater and SimplexWater (Tsur et al., 2025) provide bias-free-in-expectation constructions for low-entropy regimes, yet exhibit reduced robustness compared to standard bias-free methods under paraphrasing (AUROC of 0.87 to 0.88 under DIPPER versus 0.90 to 0.91 for DiPMark and HCW). SimMark (Dabiriaghdam & Wang, 2025) achieves performance comparable to SemStamp and PMark, confirming that sentence-level similarity watermarks remain subject to edit-rate-driven degradation.

Empirical results are consistent with theoretical predictions. Figure 2(a) shows scaling consistent with Theorem 3.2: greedy decoding exhibits $O(1)$ Pinsker-based

Table 3. Robustness and detectability on Llama-2-7B across attack conditions. For each condition, we report AUROC (AUC), TPR at 1% FPR, and the keyless z-score ((Liu et al., 2025)). Higher AUC and TPR indicate greater robustness, whereas lower z-scores indicate lower detectability under this detector. Superscripts identify watermark families: ^{B} Biased, ^{F} Bias-free, ^{S} Semantic, ^{D} Dist-preserving, ^{W} Training-time, and * Hybrid. DAWA is shown separately as a co-designed baseline. The summarization column (†) reports the results of the summarization attack from *WaterJudge* with $\hat{\varepsilon} \approx 0.55$.

	No attack			DIPPER ($\hat{\varepsilon}\approx0.25$)			OPT-2.7B ($\hat{\varepsilon}\approx0.15$)			WM-removal ($\hat{\varepsilon}\approx0.15$)			Synonym ($\hat{\varepsilon}\approx0.15$)			Back-trans. ($\hat{\varepsilon}\approx0.42$)			Summ. † ($\hat{\varepsilon}\approx0.55$)		
Method	AUC	TPR	z	AUC	TPR	z	AUC	TPR	z	AUC	TPR	z	AUC	TPR	z	AUC	TPR	z	AUC	TPR	z
KGW^{B}	.99	1.00	30.1	.86	.64	9.6	.78	.59	8.4	.78	.59	8.1	.78	.59	8.3	.58	.52	−1.2	.52	.14	−2.3
Unigram^{B}	.99	1.00	11.2	.88	.67	8.8	.79	.62	7.9	.79	.61	7.6	.79	.62	7.8	.57	.51	−1.0	.53	.15	−2.0
DiPMark^{F}	.99	1.00	43.2	.90	.80	3.9	.91	.86	3.6	.90	.85	3.4	.91	.86	3.5	.60	.54	−0.8	.55	.18	−1.1
HCW^{F}	.99	1.00	105	.91	.82	3.4	.92	.88	3.1	.92	.88	3.0	.92	.88	3.0	.59	.53	−0.6	.56	.20	−0.8
HeavyWater^{F}	.99	1.00	38.5	.88	.76	4.2	.89	.82	3.8	.88	.81	3.5	.89	.82	3.6	.56	.50	−1.0	.53	.16	−1.3
SimplexWater^{F}	.99	1.00	35.2	.87	.74	4.5	.88	.80	4.0	.87	.79	3.7	.88	.80	3.8	.55	.49	−1.1	.52	.15	−1.4
Kuditipudi^{F}	.99	1.00	27.5	.93	.85	3.4	.94	.89	3.7	.93	.88	3.5	.94	.89	3.6	.65	.43	−1.2	.62	.28	−1.6
SemStamp^{S}	.98	.98	8.5	.93	.87	2.8	.94	.91	2.4	.93	.90	2.2	.94	.91	2.3	.82	.75	1.5	.66	.38	0.5
PMark^{S}	.99	.99	7.2	.94	.89	2.5	.95	.93	2.1	.94	.92	1.9	.95	.93	2.0	.85	.79	1.2	.64	.43	0.3
SimMark^{S}	.98	.97	9.1	.92	.85	3.0	.93	.89	2.6	.92	.88	2.4	.93	.89	2.5	.80	.72	1.7	.64	.35	0.7
CGW^{D}	.99	1.00	−5.8	.50	.15	−10	.50	.31	−5.4	.50	.31	−5.6	.50	.31	−5.5	.51	.21	−6.5	.50	.16	−6.9
GaussMark^{W}	1.00	1.00	12.4	.82	.71	8.6	.85	.76	7.8	.83	.74	7.3	.84	.75	7.5	.62	.55	2.1	.56	.21	0.4
DAWA	1.00	1.00	2.2	.54	.03	1.6	.74	.05	1.3	.71	.07	1.4	.84	.08	3.0	.75	.03	2.0	.63	.02	1.0
Hybrid*	.99	1.00	−7.8	.94	.89	4.2	.96	.93	3.8	.95	.92	4.3	.95	.93	4.4	.86	.79	1.8	.61	.47	0.6

TV proxy, biased sampling grows as $\|\delta\|\sqrt{T}$, bias-free shows $\sqrt{T}$ growth, semantic schemes exhibit $O(\sqrt{T_s}) \approx O(\sqrt{T/l})$ accumulation, and distribution-preserving maintains near-zero values. Figure 2(b) illustrates detection performance under varying paraphrasing intensities, with knee points at which the Neyman-Pearson test maintains 99% power. For our models and paraphrasers, observed thresholds of $\hat{\varepsilon} \approx 0.15$ for biased and 0.32 for semantic schemes fall in the range predicted by Theorem 3.4. Figure 2(c) visualizes the empirical tradeoff among evaluated methods, where no single family achieves both high verification and low detectability across all regimes, and the hybrid consistently lies near the empirical Pareto frontier.

6. Conclusion

This work formalizes the trade-off between detectability and post-edit verification in inference-time LLM watermarking by modeling verification as hypothesis testing under output editing. We show that probability-modifying token-level schemes accumulate a persistent signal but introduce statistical drift detectable by keyless observers, whereas distribution-preserving schemes achieve keyless indistinguishability yet remain fragile to edits. Semantic schemes maintain post-edit verification by shifting evidence to sentence-level units, provided the semantic flip rate remains small relative to the token perturbation rate. The limits we derive are information-theoretic, with computational attainability depending on detector access. These insights yield a minimal-information selection principle: choose the watermark family that meets target verification with the least per-unit KL budget. The resulting hybrid selector is empirically near Pareto across evaluated edit regimes while maintaining low detectability. Our contributions provide a unified lens for comparing watermarking schemes and offer actionable guidance for practitioners.

Impact Statement

Practical Deployment. Our results reveal an inherent trade-off between robustness, statistical stealth, and reliable verification. Controlled deployments (e.g., enterprise or academic settings) can favor low-distortion designs with strong key management and access control, while public-facing deployments should prefer schemes with explicit detectability targets and regular auditing, since adaptive attacks may shift the dominant edit regime over time. Scheme choice should be conditioned on the observed edit regime: semantic watermarking is suitable for meaning-preserving paraphrases, whereas token-level methods are preferable for substantial semantic drift.

Future Work. Our analysis adopts tractable editing models parameterized by token edit rate ε and semantic flip rate $\varepsilon_s(\varepsilon)$, which capture the dominant structure of common attacks but may not fully represent sophisticated paraphrasing strategies that decouple token and semantic changes. Extending the framework to accommodate richer edit models, including watermark-aware adaptive rewriting attacks, therefore constitutes a natural next step. A related challenge is that regime-aware selection relies on estimating edit levels, which may prove difficult under adversarial conditions; addressing this limitation will require developing robust estimation methods alongside conservative selection heuristics. From a practical standpoint, deployment guidance would benefit from evaluation protocols that jointly report robustness, detectability, and output quality under realistic post-processing pipelines. Finally, training-time watermarks (Gu et al., 2024; Block et al., 2025) present a complementary research direction, particularly concerning their resistance to fine-tuning and the quality implications of model distillation.

References

Block, A., Rakhlin, A., and Sekhari, A. Gaussmark: A Practical Approach for Structural Watermarking of Language Models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/pdf?id=YG3DbpAQBf.

Christ, M., Gunn, S., and Zamir, O. Undetectable Watermarks for Language Models. In Agrawal, S. and Roth, A. (eds.), *Proceedings of Thirty-Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1125–1139. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/christ24a.html.

Dabiriaghdam, A. and Wang, L. SimMark: A Robust Sentence-Level Similarity-Based Watermarking Algorithm for Large Language Models. *arXiv preprint arXiv:2502.02787*, 2025. URL https://arxiv.org/pdf/2502.02787.

Giboulot, E. and Furon, T. Watermax: Breaking the LLM Watermark Detectability-Robustness-Quality Trade-off. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=HjeKHxK2VH.

Gloaguen, T., Jovanović, N., Staab, R., and Vechev, M. Black-box detection of language model watermarks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=E4LAVLXAHW.

Golowich, N. and Moitra, A. Edit Distance Robust Watermarks for Language Models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pp. 20645–20693, 2024. URL https://openreview.net/pdf?id=FZ45kf5pIA.

Gu, C., Li, X. L., Liang, P., and Hashimoto, T. 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.

He, H., Liu, Y., Wang, Z., Mao, Y., and Bu, Y. Theoretically Grounded Framework for LLM Watermarking: A Distribution-Adaptive Approach. In *The 1st Workshop on GenAI Watermarking*, 2025. URL https://openreview.net/forum?id=Lzi8raVEQu.

Hou, A., Zhang, J., He, T., Wang, Y., Chuang, Y.-S., Wang, H., Shen, L., Van Durme, B., Khashabi, D., and Tsvetkov, Y. SemStamp: A Semantic Watermark with Paraphrastic Robustness for Text Generation. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4067–4082, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.226. URL https://aclanthology.org/2024.naacl-long.226/.

Hu, Z., Chen, L., Wu, X., Wu, Y., Zhang, H., and Huang, H. 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.

Huo, J., Liu, S., Wang, B., Zhang, J., Yan, Y., Liu, A., Hu, X., and Zhou, M. PMark: Towards Robust and Distortion-free Semantic-level Watermarking with Channel Constraints. *arXiv preprint arXiv:2509.21057*, 2025. URL https://arxiv.org/abs/2509.21057.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A Watermark for Large Language Models. pp. 17061–17084, 2023. URL https://openreview.net/pdf?id=aX8ig9X2a7.

Kirchenbauer, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., Fernando, K., Saha, A., Goldblum, M., and Goldstein, T. 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.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., and Iyyer, M. 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.

Kuditipudi, R., Thickstun, J., Hashimoto, T., and Liang, P. Robust Distortion-free Watermarks for Language Models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=FpaCL1MO2C.

Li, X., Ruan, F., Wang, H., Long, Q., and Su, W. J. Robust detection of watermarks for large language models under human edits. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. qkaf056, 2025. URL https://doi.org/10.1093/jrsssb/qkaf056.

Liang, J., Wang, Z., Hong, S., Ji, S., and Wang, T. Watermark under Fire: A Robustness Evaluation of LLM Watermarking. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21050–21074, 2025. URL https://aclanthology.org/2025.findings-emnlp.1148/.

Liu, A., Pan, L., Hu, X., Meng, S., and Wen, L. A Semantic Invariant Robust Watermark for Large Language Models. *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/pdf?id=6p8lpe4MNf.

Liu, A., Guan, S., Liu, Y., Pan, L., Zhang, Y., Fang, L., Wen, L., Yu, P. S., and Hu, X. 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.

Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., and Harchaoui, Z. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Tqx7nJp7PR.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

Tsur, D., Long, C. X., Verdun, C. M., Vithana, S., Hsu, H., Chen, C.-F., Permuter, H. H., and Calmon, F. Heavywater and simplexwater: Distortion-free LLM watermarks for low-entropy distributions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=R5EBtNE2Y9.

Wu, Y., Hu, Z., Guo, J., Zhang, H., and Huang, H. A Resilient and Accessible Distribution-Preserving Watermark for Large Language Models. *ICML*, 2024. URL https://openreview.net/pdf?id=c8qWiNiqRY.

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

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*, 2022. URL https://arxiv.org/pdf/2205.01068.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. BERTScore: Evaluating Text Generation with BERT. *arXiv preprint arXiv:1904.09675*, 2019. URL https://arxiv.org/abs/1904.09675.

Zhao, X., Ananth, P., Li, L., and Wang, Y.-X. Provable Robust Watermarking for AI-Generated Text. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/pdf?id=SsmT8aO45L.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NeurIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc. URL https://dl.acm.org/doi/10.5555/3666122.3668142.

A. Extended Review of LLM Watermarking Literature

This appendix provides technical analysis of existing watermarking schemes for large language models, extending Section 2. We organize prior work by design principles and analyze theoretical guarantees, practical limitations, and empirical vulnerabilities.

A.1. Token-Level Probability-Modifying Watermarks

Token-level probability-modifying watermarks alter sampling distributions to embed detectable signals, differing primarily in how they manage the resulting distributional shift.

Biased Sampling. The seminal KGW scheme (Kirchenbauer et al., 2023) partitions the vocabulary at each position t into green and red lists using a hash of preceding tokens, then modifies logits via $\hat{\ell}_t[v] = \ell_t[v] + \delta \cdot \mathbf{1}[v \in G_t]$. This induces an exponential tilt increasing green token probabilities by approximately e^δ. The Unigram-Watermark (Zhao et al., 2024) fixes the partition across all positions, yielding superior robustness with tight quality bounds through Rényi divergence analysis. Both schemes accumulate outsider evidence at rate $O(|\delta|\sqrt{T})$, creating detectable frequency shifts identifiable through chi-square tests or adversarial prompting.

Bias-Free Sampling. These approaches preserve expected token distributions through reweighting functions satisfying $\mathbb{E}_E[R_E(p_t)] = p_t$ (Hu et al., 2024), vocabulary permutations (Wu et al., 2024), or inverse transform sampling (Kuditipudi et al., 2024). Despite first-order unbiasedness, all such methods introduce higher-order variance signatures growing as $O(\sqrt{T})$, detectable through second-moment tests (Gloaguen et al., 2025). HeavyWater and SimplexWater (Tsur et al., 2025) formulate watermark design as minimax optimization over distortion-free couplings, achieving improved detection in low-entropy regimes through simplex codes and heavy-tailed score distributions respectively.

A.2. Semantic-Level Watermarks

Semantic-level methods embed signals at sentence granularity to improve robustness against paraphrasing. Given context π and base distribution $P_M(\cdot \mid \pi)$, these methods select sentences satisfying a keyed predicate $A_k(s) = 1$ via rejection sampling: $P_M^w(s \mid \pi; k) = \frac{P_M(s|\pi)\,\mathbf{1}\{A_k(s)=1\}}{P_M(A_k=1|\pi)}$.

SemStamp (Hou et al., 2024) implements A_k using locality-sensitive hashing of sentence embeddings with margin constraints preventing small perturbations from flipping watermark bits. PMark (Huo et al., 2025) achieves key-averaged distortion-free selection through balanced partitions, though the distribution remains biased for any fixed key. SIR (Liu et al., 2024) encodes watermarks into meaning-preserving features surviving rewording. SimMark (Dabiriaghdam & Wang, 2025) operates without model logits by accepting sentences only when consecutive sentence embedding similarities fall within a keyed interval, using soft counting to improve paraphrase resilience. These schemes trade token-level artifacts for sentence-level selection artifacts without eliminating the stealth-robustness tension.

A.3. Alternative Embedding Strategies

Several approaches depart from direct probability modification at generation time.

Generator-Side Selection. WaterMax (Giboulot & Furon, 2024) generates K candidates from the base distribution and selects the one maximizing a keyed score: $y^* = \arg\max_{i \in [K]} \text{score}_k(y^{(i)})$. This preserves per-sample quality but induces distributional shift at the text level, with computational cost scaling linearly in K.

Detector-in-the-Loop Co-Design. DAWA (He et al., 2025) jointly optimizes the embedding rule and detector to maximize detection power under explicit false-positive and distortion constraints. The optimal scheme adapts to the underlying token distribution rather than applying uniform biases, though optimization against a particular detector may limit cross-detector generalization.

Training-Time Methods. GaussMark (Block et al., 2025) embeds watermarks by perturbing model weights with a secret Gaussian key, detecting via gradient correlation tests. Such methods remain detectable insofar as weight perturbations induce output distribution shifts.

Evaluation Standards. Benchmarking platforms such as WaterPark (Liang et al., 2025) emphasize adaptive watermark-aware attacks, quality-matched comparisons, and statistical rigor. We incorporate these principles in Section 5.

A.4. Distribution-Preserving Watermarks

Distribution-preserving schemes achieve provable undetectability by maintaining $q_t \equiv p_t$ while replacing true randomness with pseudorandom functions (Christ et al., 2024), thereby shifting the focus from statistical hypothesis testing to cryptographic verification. However, these schemes collapse under light paraphrasing since verification depends on intact PRF alignment; strengthening robustness reintroduces detectable drift.

A.5. Detection Methods

Sophisticated detection methods expose artifacts across all scheme categories. For expectation-preserving schemes, second-moment tests detect variance anomalies via $T = \sum_{t=1}^{n} \left(\|\hat{p}_t\|_2^2 - \mathbb{E}[\|p_t\|_2^2] \right)$ (Gloaguen et al., 2025). Active attacks using adversarial prompting amplify watermark biases, reducing required sample sizes by orders of magnitude (Liu et al., 2025). These vulnerabilities are structural: biased schemes are exposed via frequency analysis, bias-free schemes via variance anomalies, and both succumb to adversarial prompting.

For detection under human edits, Li et al. (2025) model edited text as a mixture where only an ε fraction of positions retain watermark evidence. They derive a phase transition: detection is possible if and only if $q + 2p < 1$, where $\varepsilon_n \asymp n^{-p}$ and token entropy scales as n^{-q}. Their Truncated Goodness-of-Fit (Tr-GoF) test achieves this optimal boundary adaptively, outperforming sum-based detectors whose boundary $q + p = 1/2$ loses substantial power under edits.

A.6. Coding-Theoretic Constructions

Golowich & Moitra (2024) construct watermarks using indexing pseudorandom codes that tolerate a constant fraction of adversarial edits in high-entropy substrings. The scheme transforms a binary pseudorandom code into an indexing PRC using symbols from a larger alphabet, with redundancy handling insertions and deletions. For entropy-rate parameter $\alpha \in (0, 1)$, the scheme tolerates edit fraction $p = \Theta(\alpha^2)$, requiring alphabet size: $|\Sigma(\lambda)| \geq n(\lambda)^{C_2 \frac{1}{\alpha} \log \frac{1}{\alpha}}$.

The polynomial degree grows as $\Theta\left(\frac{1}{\alpha} \log \frac{1}{\alpha}\right)$, implying alphabet sizes far exceeding practical LLM vocabularies of 30k to 100k tokens for realistic entropy levels. This provides an achievability result for large tunable alphabets, whereas our Theorems 3.2 and 3.4 establish information-theoretic impossibility for fixed-vocabulary LLMs. The results address complementary parameter regimes.

B. Notation and Variables

Notation Conventions

We adopt the following notational conventions throughout. $\mathcal{V}^T$ denotes T-length token sequences from vocabulary $\mathcal{V}$, with subscript t indexing token position from 1 to T. Superscripts on Q indicate the sampling method, while an asterisk $(*)$ denotes optimal values. All conditionals such as $p_t(\cdot)$ implicitly depend on the preceding context $\mathbf{y}_{<t}$ and the prompt x.

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	$\sigma^2(v) = \mathrm{Var}_E[R_E(p_t)(v)]$, $\hat{\sigma}^2 = \sum_v p_t(v)\sigma^2(v)$	§3, App. D
Z_t	$\{0, 1\}$	Keyed binary indicator: $Z_t := g_k(F(S_t))$	§3
F	Function	Semantic feature map for sentence embedding	§3
g_k	Function	Keyed predicate mapping embeddings to $\{0, 1\}$	§3
ρ	$(0, \frac{1}{2}]$	Semantic bias parameter controlling watermark strength	§3
PRF	Function	Pseudorandom function for RNG replacement	§3
$\mathcal{W}, \mathcal{W}^\star(\varepsilon)$	Scheme	Watermarking scheme and the optimal hybrid	§4, App. E
$\mathcal{H}$	Rule	Hybrid watermark selection rule	§4
K, t	Scalars	DP verifier: marked positions K and correction radius t	§4, App. E

Important Variables and Distributions

Symbol	Type/Dim	Description	Sections
L, T	Scalar	Text length (number of tokens)	§3, §4
T_s	Scalar	Number of sentences	§3, §4
$\mathcal{V}, \Sigma$	Set	Token vocabulary	§3
$\Sigma^\star$	Set	Set of all finite sequences over Σ	§3
Ω	Set	Space of complete texts	§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
S_t	$\Sigma^\star$	Sentence at generation step t	§3
$\tilde{S}_t$	$\Sigma^\star$	Edited sentence at step t	§3
$p_t(\cdot), p_\theta(\cdot\|\cdot)$	Function	Baseline LLM conditional probabilities	§3
$q_t(\cdot)$	Function	Watermarked conditional probabilities	§3
$s, \tilde{s}$	Rule	Baseline and watermarked sampling rules	§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
Q_k^{sem}	Distribution	Semantic watermarked distribution with key k	§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
$\mathrm{Bern}(p)$	Distribution	Bernoulli distribution with parameter p	§3

Detectability and Robustness

Symbol	Type/Dim	Description	Sections
$\mathrm{Detect}_{\mathrm{IT}}(\tilde{s})$	$[0, 1]$	Information-theoretic detectability	§3
$\mathrm{Detect}_{\mathrm{comp}}(\tilde{s}_\lambda)$	$[0, 1]$	Computational detectability	§3
$\mathrm{Adv}_{D_\lambda}(\lambda)$	$[0, 1]$	Distinguishing advantage for detector D_λ	§3
D_λ	Detector	Probabilistic polynomial-time detector	§3
λ	Scalar	Security parameter	§3
PPT	Set	Class of probabilistic polynomial-time algorithms	§3
$\mathrm{ED}(y, \tilde{y})$	Scalar	Edit distance between sequences	§3
$\mathrm{TV}(P, Q)$	$[0, 1]$	Total variation distance	§3, App. C
$\mathrm{KL}(Q\|\|P)$	$[0, \infty)$	Kullback-Leibler divergence (base 2 in proofs)	§3
$\mathrm{Detect}(s)$	$[0, 1]$	Distinguishability for sampling rule s	§3
$\varepsilon, \hat{\varepsilon}$	$[0, 1]$	Edit rate (true and estimated)	§3, §4
$\varepsilon_s(\varepsilon)$	$[0, 1]$	Induced semantic flip rate under token edit rate ε	§3, §4
$\varepsilon_\beta^{\mathrm{tok}}$	$[0, 1]$	Maximal tolerable edit rate for token-level schemes	§3
α, β	$[0, 1]$	Detector level and miss probability (power $= 1 - \beta$)	§4, 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

Information Budgets and Channel Capacities

Symbol	Type/Dim	Description	Sections
D_0	Bits/token	Per-token information at zero edits	§3, App. D
$D_0^{(\mathrm{tok})}$	Bits/token	Per-token information for token-level schemes	§3
$D_0^{(\mathrm{biased})}$	Bits/token	Per-token information for biased sampling	§3
$D_0^{(\mathrm{bias\text{-}free})}$	Bits/token	Per-token information for bias-free sampling	§3
$D_0^{(\mathrm{sem})}$	Bits/sentence	Per-sentence information for semantic schemes	§3
D_ε	Bits/token	Per-token information at edit rate ε	App. D
$C(\varepsilon)$	Bits	Total usable information $\approx T(1-\varepsilon)^2 D_0$	§3, App. D
$C_{\mathrm{tok}}(\varepsilon)$	Bits	Token-level channel capacity: $T(1-\varepsilon)^2 D_0^{(\mathrm{tok})}$	§3
$C_{\mathrm{sem}}(\varepsilon)$	Bits	Semantic channel capacity: $T_s(1-2\varepsilon_s)^2 D_0^{(\mathrm{sem})}$	§3
$\varepsilon_\beta(T, D_0)$	$[0, 1]$	"Knee": $1 - \sqrt{\log_2(1/\beta)/(TD_0)}$	App. D
$H(\cdot), H_2(\cdot)$	Function	Entropy, binary entropy	§3

Design Rule Parameters

Symbol	Type/Dim	Description	Sections
$D_{\mathrm{req}}(\varepsilon, T, \beta)$	Bits/token	Required information: $\log_2(1/\beta)/T(1-\varepsilon)^2$	§4.1, App. D
$D^{\mathrm{tok}}_{\mathrm{req}}(\varepsilon, T, \beta)$	Bits/token	Required per-token information for token-level schemes	§4
$D^{\mathrm{sem}}_{\mathrm{req}}(\varepsilon, T_s, \beta)$	Bits/sentence	Required per-sentence information for semantic schemes	§4
M, τ	Scalar, $[0, 1]$	Outsider pooled tokens M and TV budget τ	§4.1, App. D
M_s, τ_s	Scalar, $[0, 1]$	Outsider pooled sentences M_s and TV budget τ_s	§4
$D_{\mathrm{stealth}}(M, \tau)$	Bits/token	Stealth cap $\frac{2\tau^2}{M \ln 2}$	§4.1, App. D
$D^{\mathrm{tok}}_{\mathrm{stealth}}(M, \tau)$	Bits/token	Token-level stealth cap	§4
$D^{\mathrm{sem}}_{\mathrm{stealth}}(M_s, \tau_s)$	Bits/sentence	Semantic stealth cap	§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
$\lambda_r, \lambda_q, \lambda_a$	Scalars	Weights for reliability, stealth penalty, amplitude	§4.1
$D^\star$	Bits/token	Target per-token information after constraints	§4.1, App. E
$D^{\max}_{\mathrm{BF}}, D^{\max}_{\mathrm{B}}$	Bits/token	Available budgets for BF and B families	§4.1, 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., $\sqrt{\tilde{\sigma}^2}$ 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

C. Proof of Theorem 3.2

This appendix establishes the detectability bounds stated in Theorem 3.2. The proof relies on the KL chain rule for autoregressive distributions and Pinsker's inequality relating KL divergence to total variation distance.

C.1. Preliminaries

The total variation distance and KL divergence between distributions P and Q are $\mathrm{TV}(P, Q) = \frac{1}{2}\sum_{\mathbf{y}} |P(\mathbf{y}) - Q(\mathbf{y})|$ and $\mathrm{KL}(Q\|P) = \mathbb{E}_Q[\log(Q(\mathbf{y})/P(\mathbf{y}))]$, respectively. Pinsker's inequality states $\mathrm{TV}(P, Q) \le \sqrt{\frac{1}{2}\mathrm{KL}(Q\|P)}$. For autoregressive distributions factorizing as $Q(\mathbf{y}) = \prod_{t=1}^{T} q_t(y_t \mid y_{<t})$, the KL divergence decomposes via the chain rule:

$$\mathrm{KL}(Q\|P) = \sum_{t=1}^{T} \mathbb{E}_{y_{<t}\sim Q}[\mathrm{KL}(q_t(\cdot \mid y_{<t})\|p_t(\cdot \mid y_{<t}))]\,. \tag{7}$$

This identity applies equally to token sequences (length T) and sentence sequences (length T_s). Throughout, $\log$ denotes the natural logarithm; to convert to bits, note that $\mathrm{KL}_2(\cdot\|\cdot) = \mathrm{KL}(\cdot\|\cdot)/\ln 2$.

When the main paper reports total variation numerically in experiments, it is computed via the Pinsker upper bound derived from measured KL divergence rather than through direct estimation of total variation over the full text distribution.

C.2. Greedy Sampling

Let Q^{greedy} place unit mass on the greedy path $\mathbf{y}^\star$, where $y_t^\star = \arg\max_v p_t(v \mid y^\star_{<t})$. Since $Q^{\mathrm{greedy}}(\mathbf{y}^\star) = 1$ and $Q^{\mathrm{greedy}}(\mathbf{y}) = 0$ for $\mathbf{y} \neq \mathbf{y}^\star$, direct computation yields $\mathrm{TV}(P^s, Q^{\mathrm{greedy}}) = \frac{1}{2}(1 - P^s(\mathbf{y}^\star) + 1 - P^s(\mathbf{y}^\star)) = 1 - P^s(\mathbf{y}^\star)$.

C.3. Biased Sampling

At position t, let $G_t \subseteq \mathcal{V}$ denote the keyed green set with baseline mass $g_t := p_t(G_t)$. The biased sampler applies an exponential tilt $q_t(v) = p_t(v)\exp\{\delta \mathbf{1}[v \in G_t]\}/Z_t$, where $Z_t = (1 - g_t) + g_t e^\delta$. The one-step KL divergence is

$$\mathrm{KL}(q_t\|p_t) = \delta\, q_t(G_t) - \log Z_t = \frac{g_t e^\delta}{(1-g_t) + g_t e^\delta}\,\delta - \log\left((1-g_t) + g_t e^\delta\right). \tag{8}$$

For $|\delta| \ll 1$, Taylor expansion gives $Z_t = 1 + g_t(\delta + \delta^2/2) + O(\delta^3)$, $\log Z_t = g_t\delta + g_t(1-g_t)\delta^2/2 + O(\delta^3)$, and $q_t(G_t) = g_t + g_t(1-g_t)\delta + O(\delta^2)$. Substituting yields

$$\mathrm{KL}(q_t \| p_t) = \frac{g_t(1-g_t)}{2}\delta^2 + O(\delta^3). \tag{9}$$

Applying the chain rule (7) and Pinsker's inequality, we obtain $\mathrm{TV}(P^s, Q^{\mathrm{bias}_\delta}) \le |\delta|\sqrt{\frac{1}{4}\sum_{t=1}^{T}\mathbb{E}[g_t(1-g_t)]} = O(|\delta|\sqrt{T})$.

C.4. Bias-free Sampling

A bias-free watermark applies a keyed reweighting operator R_E to the baseline distribution, producing $q_{t,E}(\cdot \mid y_{<t}) := R_E(p_t(\cdot \mid y_{<t}))$ with $Q_E^{\mathrm{bf}}(y_{1:T}) = \prod_{t=1}^{T} q_{t,E}(y_t \mid y_{<t})$. Unbiasedness requires $\mathbb{E}_E[R_E(p)] = p$ for every conditional law p. Writing the pointwise perturbation as $q_{t,E}(v) = p_t(v) + \epsilon_{t,E}(v)$ with $\sum_v \epsilon_{t,E}(v) = 0$, a second-order Taylor expansion in the small-signal regime ($|\epsilon_{t,E}(v)| \ll p_t(v)$) yields

$$\mathrm{KL}(q_{t,E} \| p_t) = \frac{1}{2}\sum_v \frac{\epsilon_{t,E}(v)^2}{p_t(v)} + O(\|\epsilon_{t,E}\|_\infty^3). \tag{10}$$

Applying the chain rule and Pinsker's inequality gives the fixed-key bound $\mathrm{TV}(P^s, Q_E^{\mathrm{bf}}) \le \sqrt{\frac{1}{4}\sum_{t=1}^{T}\mathbb{E}[\sum_v \epsilon_{t,E}(v)^2/p_t(v)]}$.

Typical-key surrogate. The main-text table reports a variance-controlled bound obtained by averaging over keys. By unbiasedness, $\mathbb{E}_E[\epsilon_{t,E}(v)] = 0$, so $\mathbb{E}_E[\epsilon_{t,E}(v)^2] = \mathrm{Var}_E[R_E(p_t)(v)]$. Taking the expectation of (10) over E gives $\mathbb{E}_E[\mathrm{KL}(q_{t,E}\|p_t)] \approx \frac{1}{2}\sum_v \mathrm{Var}_E[R_E(p_t)(v)]/p_t(v)$. Consequently,

$$\mathbb{E}_E[\mathrm{TV}(P^s, Q_E^{\mathrm{bf}})^2] \le \frac{1}{4}\sum_{t=1}^{T}\mathbb{E}\Big[\sum_v \frac{\mathrm{Var}_E[R_E(p_t)(v)]}{p_t(v)}\Big]. \tag{11}$$

By Markov's inequality, for any $\eta \in (0,1)$, the bound $\mathrm{TV}(P^s, Q_E^{\mathrm{bf}}) \le \eta^{-1/2}\sqrt{\frac{1}{4}\sum_{t,v}\mathrm{Var}_E[R_E(p_t)(v)]/p_t(v)}$ holds for a $(1-\eta)$ fraction of keys. This establishes the $O(\sqrt{T})$ scaling on average and for typical keys.

C.5. Distribution-Preserving Sampling

If a keyed pseudorandom source replaces the randomness while preserving per-step probabilities ($q_t \equiv p_t$ for all histories), then $Q^{\mathrm{prf}}(\mathbf{y}) = \prod_{t=1}^{T} p_t(y_t \mid y_{<t}) = P^s(\mathbf{y})$, giving $\mathrm{TV}(P^s, Q^{\mathrm{prf}}) = 0$.

C.6. Semantic Rejection Sampling

For sentence-level watermarks, let $\mathbf{s}_{1:T_s} = (s_1, \ldots, s_{T_s})$ with each $s_t \in \Sigma^\star$. At step t, define a keyed accept set $A_{t,k}(\mathbf{s}_{<t}) \subseteq \Sigma^\star$ with acceptance mass $a_t := p_t(A_{t,k}(\mathbf{s}_{<t}) \mid \mathbf{s}_{<t}) \in (0,1]$. Rejection sampling draws candidates from p_t until one falls in $A_{t,k}$, yielding $q_{t,k}(s \mid \mathbf{s}_{<t}) = p_t(s \mid \mathbf{s}_{<t})\mathbf{1}\{s \in A_{t,k}\}/a_t$.

For $s \in A_{t,k}$, the likelihood ratio equals $1/a_t$, so $\mathrm{KL}(q_{t,k}\|p_t) = \sum_{s\in A_{t,k}}(p_t(s)/a_t)\log(1/a_t) = \log(1/a_t)$. Note that a_t also controls sampling efficiency, as rejection sampling requires an expected $1/a_t$ draws; practical schemes enforce a_t bounded away from zero.

Applying the chain rule at the sentence level and Pinsker's inequality gives

$$\mathrm{TV}(P^s, Q_k^{\mathrm{sem}}) \le \sqrt{\frac{1}{2}\sum_{t=1}^{T_s}\mathbb{E}[\log(1/a_t)]}. \tag{12}$$

Assuming uniform acceptance $a_t \ge a_{\min} > 0$, this yields $\mathrm{TV}(P^s, Q_k^{\mathrm{sem}}) \le \sqrt{(T_s/2)\log(1/a_{\min})}$. Converting to token length via $T \approx \ell T_s$ (where ℓ is the typical sentence length) gives the $O(\sqrt{T/\ell})$ scaling.

Remark C.1 (Random quantities). The quantities g_t, $\mathrm{Var}_E[R_E(p_t)(v)]$, and a_t depend on random histories. Throughout, these appear inside expectations, so the final bounds are deterministic functions of the prompt, length, and watermark parameters.

C.7. Information-theoretic versus Computational Detectability

This subsection clarifies the relationship between information-theoretic detectability (total variation) and computational detectability (PPT distinguishers), organized according to the three detector access models defined in Section 3.

Lemma C.2 (Computational detectability is upper bounded by TV)**.** *For every pair of distributions* (P_λ, Q_λ) *over* Ω*,* $\sup_{D_\lambda \in \mathsf{PPT}} |\Pr_{Q_\lambda}[D_\lambda(y) = 1] - \Pr_{P_\lambda}[D_\lambda(y) = 1]| \le \mathrm{TV}(P_\lambda, Q_\lambda)$.

Proof. For any randomized PPT detector D_λ with internal randomness R, let $A_r := \{y : D_{\lambda,r}(y) = 1\}$ denote the acceptance set when $R = r$. For each fixed r, $|Q_\lambda(A_r) - P_\lambda(A_r)| \le \mathrm{TV}(P_\lambda, Q_\lambda)$. Averaging over R and applying Jensen's inequality yields the claim. □

Keyless versus key-holding distinguishers. For a keyed sampler with key space $\mathcal{K}_\lambda$, let $Q_{\lambda,k}$ denote the distribution induced by key k. For a *fixed key*, the optimal unbounded distinguisher achieves advantage $\mathrm{TV}(P_\lambda, Q_{\lambda,k})$. When the key is *hidden*, the keyless observation under H_1 is drawn from the mixture $\overline{Q}_\lambda := \mathbb{E}_k[Q_{\lambda,k}]$, and the optimal keyless advantage is $\mathrm{TV}(P_\lambda, \overline{Q}_\lambda)$. These notions differ materially for expectation-preserving schemes: $\overline{Q}_\lambda = P_\lambda$ (perfect key-averaged stealth) is possible even when $Q_{\lambda,k} \neq P_\lambda$ for every fixed k.

Oracle and surrogate detectors: attainability of the Neyman-Pearson bound. Under oracle access, where the detector can evaluate the true conditional probabilities $p_t(\cdot \mid x, y_{<t})$ and has access to the watermark rule (and the key, if acting as verifier), the likelihood ratio $L_{\lambda,k}(y) := Q_{\lambda,k}(y)/P_\lambda(y)$ is computable in polynomial time for non-cryptographic watermarks. Consequently, the Neyman-Pearson optimal test $D^\star_{\lambda,k}(y) := \mathbf{1}\{L_{\lambda,k}(y) \ge 1\}$ can be implemented efficiently, attaining the full TV advantage.

For greedy sampling and token-level biased or bias-free samplers, the likelihood ratio factorizes as $\log L_{\lambda,k}(y_{1:T}) = \sum_{t=1}^{T} \log(q_{t,k}(y_t \mid y_{<t})/p_t(y_t \mid y_{<t}))$ and is efficiently computable given oracle access to p_t and the key. Under surrogate access, where the detector uses an independently trained model $\hat{p}_t$ approximating the true conditionals, the same procedure applies with surrogate probabilities substituted for the true ones. The resulting test achieves power close to the Neyman-Pearson bound when the surrogate model is sufficiently accurate.

For semantic rejection-sampling schemes, the per-step ratio involves acceptance masses a_t that may not be efficiently computable when $A_{t,k}$ is a complex semantic region. This represents a non-cryptographic source of gap between information-theoretic and PPT-based detection even under oracle or surrogate access.

Sample-only detectors: upper bounds without attainability guarantees. Under sample-only access, where the detector observes only text samples without the ability to evaluate token-level probabilities from any model, the Pinsker-based bounds derived throughout this appendix remain valid as upper bounds on total variation. However, no general attainability claim can be made: the Neyman-Pearson test is not implementable without probability access, and the actual distinguishing advantage achievable by sample-only detectors may be strictly smaller than the information-theoretic bound. This limitation applies to all watermarking families.

Summary by access model. The relationship between computational and information-theoretic detectability depends on detector access:

- **Oracle or surrogate access with key (verifier setting):** For token-level biased and bias-free watermarks, the likelihood ratio is computable in polynomial time, so $\mathrm{Detect}_{\mathrm{comp}} = \mathrm{Detect}_{\mathrm{IT}}$.
- **Oracle or surrogate access without key (keyless outsider):** Keyless PPT detectors can exploit key-independent artifacts (frequency shifts for biased schemes, variance anomalies for bias-free schemes), consistent with empirical black-box detection results, though the attainable advantage may be smaller than the fixed-key TV.
- **Sample-only access:** Only upper bounds are established; no general attainability claim holds.

Cryptographic separations. Cryptographic pseudorandomness can yield distributions that are statistically far apart yet computationally indistinguishable for keyless PPT adversaries. For distribution-preserving PRF-seeded watermarks, the ideal target is $Q_{\lambda,k} = P_\lambda$, yielding $\mathrm{TV} = 0$ and hence zero detectability even information-theoretically.

D. Proof of Theorem 3.4

This appendix derives Theorem 3.4, covering both token-level probability-modifying watermarks (biased and bias-free) and semantic/sentence-level watermarks (PMark, SemStamp). Throughout, all logarithms are base 2 and KL divergences are measured in bits.

The proof follows a unified template: (i) define post-attack distributions under H_0 (unwatermarked) and H_1 (watermarked), (ii) compute per-unit KL information at zero attack, (iii) show how attacks contract this information, and (iv) aggregate across units and apply a Stein-type condition to obtain power guarantees.

D.1. Edit Channel Model

Let Σ denote the vocabulary. At each token position t, the edited token $\tilde{Y}_t$ is drawn as $\tilde{Y}_t = Y_t$ with probability $1-\varepsilon$ and $\tilde{Y}_t = U_t \sim R(\cdot)$ with probability ε, where R is a replacement distribution independent of everything else. Equivalently, for distribution P on Σ, the mixture edit channel acts as

$$T_{\varepsilon,R}(P) := (1-\varepsilon)P + \varepsilon R. \tag{13}$$

We assume R has full support with $\min_v R(v) \geq r_{\min} > 0$. The pre-noise conditionals p_t (baseline) and q_t (watermarked) are mapped to $p_{t,\varepsilon} = T_{\varepsilon,R}(p_t)$ and $q_{t,\varepsilon} = T_{\varepsilon,R}(q_t)$.

D.2. Semantic-level Abstraction

For semantic watermarks, evidence is computed per sentence even though attackers edit tokens. Fix a segmentation into T_s sentences $\mathbf{S}_{1:T_s}$. A semantic watermark defines a keyed evidence function $Z_t := g_k(F(S_t)) \in \{0,1\}$, where F is a semantic feature map and g_k is a keyed predicate. Balanced schemes satisfy $\mathbb{E}[Z_t \mid H_0] \approx \frac{1}{2}$ under the baseline, while watermarking induces bias $\mathbb{E}[Z_t \mid H_1] = \frac{1}{2} + \rho$ with $|\rho| \ll 1$.

Let $\widetilde{S}_t$ denote the sentence after token-level edits and $\widetilde{Z}_t := g_k(F(\widetilde{S}_t))$. The induced semantic flip probability is $\varepsilon_s(\varepsilon) := \Pr[\widetilde{Z}_t \neq Z_t]$, which depends on attack type and semantic feature stability. We model evidence flips as a binary symmetric channel: $\widetilde{Z}_t = Z_t \oplus N_t$ with $N_t \sim \mathrm{Bern}(\varepsilon_s(\varepsilon))$ independent across t.

D.3. Preliminaries: KL Expansions and Reliability Bound

Lemma D.1 (Second-order KL expansion)**.** *Let p be a distribution on a finite set and $q = p + r$ with $\sum_v r(v) = 0$ and $|r(v)| \leq \eta\, p(v)$ for $\eta \ll 1$. Then*

$$D(q\|p) = \frac{1}{2\ln 2} \sum_v \frac{r(v)^2}{p(v)} \cdot (1 + O(\eta)). \tag{14}$$

Proof. Using $\log(1+x) = x - x^2/2 + O(x^3)$ with $x_v = r(v)/p(v)$ and noting $\sum_v r(v) = 0$, the linear terms cancel and the quadratic terms yield the stated expression. The remainder is $O(\eta)$ times the quadratic term. □

Lemma D.2 (Stein's sufficient condition)**.** *For a binary hypothesis test between product distributions on sequences of length L, if the total KL divergence satisfies $\sum_{t=1}^{L} D(P_t^{(1)}\|P_t^{(0)}) \geq \log_2(1/\beta) + o(L)$, then for sufficiently large L the Neyman-Pearson test at level α achieves miss probability at most β.*

Proof. Let $S_L = \sum_{t=1}^{L} \log_2(P_t^{(1)}(Y_t)/P_t^{(0)}(Y_t))$ be the log-likelihood ratio. The NP test rejects H_0 when $S_L \geq \tau_L$. Under H_0, Markov's inequality with $s=1$ gives $\Pr_0(S_L \geq \tau_L) \leq 2^{-\tau_L}$; choosing $\tau_L = \log_2(1/\alpha)$ ensures level α.

Under H_1, define $\psi_t(s) := -\log_2 \sum_y P_t^{(1)}(y)^{1-s} P_t^{(0)}(y)^s$. By smoothness, $\psi_t(0) = 0$ and $\psi_t'(0) = D(P_t^{(1)}\|P_t^{(0)})$. A Chernoff bound yields $\Pr_1(S_L \leq \tau_L) \leq 2^{-(\sum_t D(P_t^{(1)}\|P_t^{(0)}) - \log_2(1/\alpha) - o(L))}$, establishing the claim. □

D.4. Per-unit Information at $\varepsilon = 0$

Token-level. For biased sampling with tilt δ toward green set G with baseline mass $\gamma = p_t(G)$, Taylor expansion of the tilted distribution $q_{t,\delta}(v) \propto p_t(v) e^{\delta \mathbf{1}[v \in G]}$ yields $D(q_{t,\delta}\|p_t) = \delta^2 \gamma(1-\gamma)/(2\ln 2) + O(\delta^3)$. For bias-free sampling with

$q_{t,E}(v) = p_t(v)(1 + \Delta_E(v))$ where $\mathbb{E}_E[\Delta_E(v)] = 0$, unbiasedness eliminates linear terms, giving $\mathbb{E}_E[D(q_{t,E}\|p_t)] = \hat{\sigma}^2/(2\ln 2) + O(\|\Delta_E\|_\infty^3)$, where $\hat{\sigma}^2 = \sum_v p_t(v)\mathrm{Var}_E[\Delta_E(v)]$.

Semantic. For Bernoulli evidence with $Z \sim \mathrm{Bern}(\frac{1}{2} + \rho)$ under H_1 and $Z \sim \mathrm{Bern}(\frac{1}{2})$ under H_0:

$$D_0^{(\mathrm{sem})} = (\tfrac{1}{2} + \rho)\log_2(1 + 2\rho) + (\tfrac{1}{2} - \rho)\log_2(1 - 2\rho) = \frac{2\rho^2}{\ln 2} + O(\rho^4). \tag{15}$$

The quadratic form follows from Taylor expanding $\log_2(1 \pm 2\rho)$; odd-order terms cancel by symmetry.

D.5. Edits Contract the Signal Quadratically

Lemma D.3 (Mixture-channel KL contraction)**.** *Let $q = p + r$ with $|r(v)| \le \eta\, p(v)$ for $\eta \ll 1$, and let $p_\varepsilon = T_{\varepsilon,R}(p)$, $q_\varepsilon = T_{\varepsilon,R}(q)$. Then*

$$D(q_\varepsilon\|p_\varepsilon) = (1 + o(1))(1 - \varepsilon)^2 D(q\|p). \tag{16}$$

Proof. The perturbation under editing is $\widetilde{r} = q_\varepsilon - p_\varepsilon = (1 - \varepsilon)r$. By Lemma D.1, $D(q_\varepsilon\|p_\varepsilon) = (1/(2\ln 2))\sum_v \widetilde{r}(v)^2/p_\varepsilon(v) + O(\cdot)$. Substituting $\widetilde{r}(v) = (1-\varepsilon)r(v)$ and expanding $1/p_\varepsilon(v) = 1/((1-\varepsilon)p(v) + \varepsilon R(v)) = (1/p(v))(1 + O(\varepsilon))$ yields the $(1 - \varepsilon)^2$ factor. □

This $(1-\varepsilon)^2$ contraction holds for any mixture channel that linearly attenuates the perturbation, not just uniform substitution. For correlated paraphrasing attacks, empirical observations suggest similar quadratic attenuation when ε is calibrated to observed token edit rates.

Lemma D.4 (Semantic evidence contraction)**.** *Under the BSC model $\widetilde{Z} = Z \oplus N$ with $N \sim \mathrm{Bern}(\varepsilon_s)$:*

(i) Under H_0, $\widetilde{Z} \sim \mathrm{Bern}(\frac{1}{2})$ (unbiased bits are invariant under symmetric flips).
(ii) Under H_1, $\widetilde{Z} \sim \mathrm{Bern}(\frac{1}{2} + \rho_\varepsilon)$ with $\rho_\varepsilon = (1 - 2\varepsilon_s)\rho$.
(iii) The post-attack KL satisfies $D_\varepsilon^{(\mathrm{sem})} = (1 - 2\varepsilon_s)^2 D_0^{(\mathrm{sem})} + O(\rho^4)$.

Proof. Part (i): $\Pr(\widetilde{Z} = 1) = \frac{1}{2}(1-\varepsilon_s) + \frac{1}{2}\varepsilon_s = \frac{1}{2}$. Part (ii): $\Pr(\widetilde{Z} = 1) = (\frac{1}{2} + \rho)(1-\varepsilon_s) + (\frac{1}{2} - \rho)\varepsilon_s = \frac{1}{2} + \rho(1 - 2\varepsilon_s)$. Part (iii) follows from (15) with ρ replaced by ρ_ε. □

D.6. Sequence-level Aggregation

By the KL chain rule, the total divergence decomposes as $D(Q_\varepsilon\|P_\varepsilon) = \sum_{t=1}^T \mathbb{E}_{\tilde{Y}_{<t}\sim Q_\varepsilon}[D(Q_\varepsilon(\tilde{Y}_t \mid \tilde{Y}_{<t})\|P_\varepsilon(\tilde{Y}_t \mid \tilde{Y}_{<t}))]$. In the small-signal regime, each conditional inherits quadratic attenuation, yielding

$$C_{\mathrm{tok}}(\varepsilon) := D(Q_\varepsilon\|P_\varepsilon) \approx T(1 - \varepsilon)^2 D_0 \quad \text{(bits)}. \tag{17}$$

For semantic evidence under the independence approximation, the i.i.d. structure gives exact additivity:

$$C_{\mathrm{sem}}(\varepsilon) := D(P^{(1)}_{\widetilde{Z}_{1:T_s}}\|P^{(0)}_{\widetilde{Z}_{1:T_s}}) = T_s D_\varepsilon^{(\mathrm{sem})} \approx T_s(1 - 2\varepsilon_s(\varepsilon))^2 D_0^{(\mathrm{sem})}. \tag{18}$$

D.7. Power Condition and Knee Edit Rate

Applying Lemma D.2 with $C(\varepsilon) \ge \log_2(1/\beta)$ yields the knee, which is the maximal edit rate compatible with target power.

Token-level: From $T(1 - \varepsilon)^2 D_0 \ge \log_2(1/\beta)$, solving for ε gives

$$\varepsilon_\beta(T, D_0) = 1 - \sqrt{\frac{\log_2(1/\beta)}{TD_0}}. \tag{19}$$

Semantic-level: From $T_s(1 - 2\varepsilon_s)^2 D_0^{(\mathrm{sem})} \ge \log_2(1/\beta)$, solving for ε_s gives

$$\varepsilon_s(\varepsilon) \le \frac{1}{2}\left(1 - \sqrt{\frac{\log_2(1/\beta)}{T_s D_0^{(\mathrm{sem})}}}\right). \tag{20}$$

D.8. Impossibility Region

The Chernoff-Stein converse states that under regularity conditions, the optimal miss probability satisfies $-\log_2 \beta_L^\star(\alpha) = D(Q^{(L)}\|P^{(L)}) + o(L)$. Thus, if $C(\varepsilon) \le \log_2(1/\beta) - \omega(1)$, no level-$\alpha$ test can achieve miss probability β for large L. Combined with Lemma D.2, this yields asymptotically tight boundaries: $T(1-\varepsilon)^2 D_0 \gtrsim \log_2(1/\beta)$ for token-level and $T_s(1-2\varepsilon_s)^2 D_0^{(\mathrm{sem})} \gtrsim \log_2(1/\beta)$ for semantic schemes.

D.9. Scope of Validity

Token-level statements require the small-signal regime ($|\delta| \ll 1$ for biased, $\|\Delta_E\|_\infty \ll 1$ for bias-free) with $p_t(v)$ bounded away from zero. Lemma D.1 quantifies approximation error as lower-order. Semantic statements require $|\rho| \ll 1$ and the BSC abstraction where attack effects are summarized by $\varepsilon_s(\varepsilon)$. The mapping $\varepsilon \mapsto \varepsilon_s(\varepsilon)$ is attack-dependent and estimated empirically.

D.10. Conclusion of Proof

Combining per-unit KL expressions, quadratic attenuation under edits, chain rule aggregation, and Stein's condition yields:

$$\text{Token-level:} \quad T(1-\varepsilon)^2 D_0 \ge \log_2(1/\beta) \Rightarrow \varepsilon_\beta = 1 - \sqrt{\log_2(1/\beta)/(TD_0)}, \tag{21}$$

$$\text{Semantic:} \quad T_s(1-2\varepsilon_s)^2 D_0^{(\mathrm{sem})} \ge \log_2(1/\beta) \Rightarrow \varepsilon_s \le \tfrac{1}{2}(1 - \sqrt{\log_2(1/\beta)/(T_s D_0^{(\mathrm{sem})})}). \tag{22}$$

□

Proof of Corollary 3.5

Theorem 3.4 requires $T(1-\varepsilon)^2 D_0 \ge \log_2(1/\beta)$ for power $1-\beta$. For $\varepsilon > \varepsilon_\beta(T, D_0)$, this inequality is violated and reliable detection is unattainable.

For stealth-aware tightening, suppose outsiders may pool M tokens with TV constraint τ. Pinsker's inequality implies $D_0 \le (2/\ln 2)\tau^2/M$. Substituting yields $\varepsilon \le 1 - \sqrt{(\log_2(1/\beta)/T) \cdot (M \ln 2)/(2\tau^2)}$; edit rates exceeding this are infeasible under the stealth constraint. □

D.11. Information-theoretic vs. Computational Hardness

Lemma D.5 (Computational power bounded by IT power)**.** *For any λ, ε, and α:* $\mathrm{Power}_{\mathrm{comp},\lambda}(\varepsilon,\alpha) \le \mathrm{Power}_{\mathrm{IT},\lambda}(\varepsilon,\alpha)$.

Proof. The Neyman-Pearson test maximizes power among all level-α tests. Since PPT detectors form a subset of all detectors, the supremum over PPT cannot exceed the information-theoretic supremum. □

Equality for non-cryptographic families under oracle or surrogate access. For biased and bias-free watermarks, the post-edit likelihood ratio $\log(Q_\varepsilon(y_{1:T})/P_\varepsilon(y_{1:T})) = \sum_{t=1}^T \log(q_{t,\varepsilon}(y_t)/p_{t,\varepsilon}(y_t))$ is computable in $O(T)$ time given oracle or surrogate access to model probabilities and watermark parameters. The NP test thus runs in PPT and attains the IT power boundary, so $\mathrm{Power}_{\mathrm{comp}} = \mathrm{Power}_{\mathrm{IT}}$ for these families under oracle or surrogate access.

Sample-only access. Under sample-only access, where the detector cannot evaluate token-level probabilities, the NP test is not implementable. The Pinsker-based upper bounds on total variation remain valid, but no attainability claim is made for sample-only detectors.

Semantic schemes may exhibit a gap. Evaluating semantic likelihood ratios requires computing acceptance masses $a_t = p_t(A_{t,k} \mid \mathbf{s}_{<t})$, which may involve intractable normalization over semantic regions. Thus semantic schemes may have $\mathrm{Power}_{\mathrm{comp}} < \mathrm{Power}_{\mathrm{IT}}$ even without cryptographic assumptions.

Cryptographic separations. Consider a PRG $G : \{0,1\}^\lambda \to \{0,1\}^{T(\lambda)}$ with $P_\lambda = U_{T(\lambda)}$ and $Q_\lambda = \mathsf{Law}(G(U_\lambda))$. The sequence-level KL is $D(Q_\lambda\|P_\lambda) = T(\lambda) - \lambda$, so after editing, $C_{\mathrm{IT}}(\lambda,\varepsilon) \approx (1-\varepsilon)^2(T(\lambda)-\lambda)$ grows with $T(\lambda)$. An IT detector achieves high power for large $T(\lambda)$.

However, any PPT detector with non-negligible power would yield a PRG distinguisher: given z from the PRG game, sample $\tilde{z} \sim T_\varepsilon(\delta_z)$ and output the detector's guess. Since $\tilde{z} \sim P_{\lambda,\varepsilon}$ when z is uniform and $\tilde{z} \sim Q_{\lambda,\varepsilon}$ when z is pseudorandom,

non-negligible detection power contradicts PRG security. Thus $\mathrm{Power}_{\mathrm{comp}} \le \mathrm{negl}(\lambda)$ even when $\mathrm{Power}_{\mathrm{IT}} \ge 1-\beta$.

Empirical support. Black-box detectors (Gloaguen et al., 2025) achieve near-perfect AUROC when $T_{\mathrm{tot}} D_0 \gtrsim \log(1/\beta)$, consistent with our bounds. Their methods exploit frequency shifts (biased) and variance anomalies (bias-free) without internal model access, supporting the conclusion that no meaningful statistical-computational gap exists for non-cryptographic families under surrogate access.

D.12. Relation to Coding-theoretic Bounds

The scaling $C(\varepsilon) \approx T(1-\varepsilon)^2 D_0$ should not be interpreted as a capacity formula for generic edit channels. Classical insertion-deletion codes achieve rate $1 - O(\varepsilon)$ when encoders freely select arbitrary codewords. Our setting differs fundamentally: all outputs must lie in the typical set of a fixed base model P, a stealth constraint limits per-token KL drift D_0, and the task is binary hypothesis testing rather than message recovery. Theorem 3.4 translates the Chernoff-Stein criterion into an explicit stealth-constrained robustness bound for distribution-constrained watermarking.

E. Proof Details for Section 4

This appendix provides derivations supporting Section 4, organized as follows: the KL-based reliability condition; token-level and semantic-level information expressions; stealth caps via Pinsker's inequality; the minimal-information optimality theorem; the complete proof of the family-selection rule; and the estimation protocol for $(\varepsilon, \varepsilon_s(\varepsilon))$ with sensitivity analysis. Throughout, all logarithms are base 2, so that KL divergences are measured in bits.

E.1. Reliability via KL Divergence

The backbone of Section 4 is the standard Chernoff-Stein sufficiency rule: if the sequence-level KL divergence under H_1 relative to H_0 exceeds $\log_2(1/\beta)$ up to lower-order terms, then there exists a level-α Neyman-Pearson test with miss probability at most β.

Lemma E.1 (KL sufficiency for miss probability β)**.** *Consider a binary hypothesis test between distributions (P, Q) with false-alarm constraint $\Pr_{Y\sim P}[D(Y) = 1] \le \alpha$. Suppose that the log-likelihood ratio decomposes as a sum of independent contributions with a finite moment-generating function near the origin. Then, for any fixed $\alpha, \beta \in (0, 1)$ and sufficiently large blocklengths, there exists a level-α test with miss probability at most β whenever*

$$\sum_{t=1}^{L} D(P_t^{(1)} \| P_t^{(0)}) \;\ge\; \log_2 \frac{1}{\beta} + \log_2 \frac{1}{\alpha} + o(L). \tag{23}$$

For fixed α, this condition becomes $D(Q\|P) \ge \log_2(1/\beta) + O(1)$, motivating the leading-order threshold $\log_2(1/\beta)$ used in the main text.

E.2. Token-Level Channel Derivation

Let $Y_{1:T}$ denote the generated token sequence. Under H_0, the baseline autoregressive distribution is $P^s(y_{1:T}) = \prod_{t=1}^{T} p_t(y_t \mid y_{<t})$, whereas under a token-level watermark with a fixed key, the pre-edit distribution is $Q(y_{1:T}) = \prod_{t=1}^{T} q_t(y_t \mid y_{<t})$.

The attacker applies a token substitution channel at rate ε, mapping each conditional distribution via $T_{\varepsilon,R}(P) := (1-\varepsilon)P + \varepsilon R$, where R denotes a fixed replacement distribution with full support. In the small-signal regime, writing $q_t = p_t + r_t$ with $\sum_v r_t(v) = 0$, the linearity of $T_{\varepsilon,R}$ yields $q_{t,\varepsilon} - p_{t,\varepsilon} = (1-\varepsilon) r_t$. Since KL divergence is locally quadratic in the perturbation (Lemma D.1), the per-token KL contracts by $(1-\varepsilon)^2$ to second order:

$$D(q_{t,\varepsilon} \| p_{t,\varepsilon}) = (1 + o(1))(1-\varepsilon)^2 D(q_t \| p_t). \tag{24}$$

Aggregating across tokens via the chain rule gives $C_{\mathrm{tok}}(\varepsilon) := D(Q_\varepsilon \| P_\varepsilon^s) \approx T(1-\varepsilon)^2 D_0^{(\mathrm{tok})}$. Applying Lemma E.1 yields the sufficient condition $T(1-\varepsilon)^2 D_0^{(\mathrm{tok})} \ge \log_2(1/\beta)$, which rearranges to $D_0^{(\mathrm{tok})} \ge D_{\mathrm{req}}^{\mathrm{tok}}(\varepsilon, T, \beta) := \log_2(1/\beta)/[T(1-\varepsilon)^2]$, matching (1).

E.3. Semantic Evidence Model and Contraction

Let the generated text be segmented into T_s sentences. A semantic watermark induces, for each sentence i, a keyed evidence statistic $Z_i \in \{0, 1\}$. We adopt the small-signal mean-shift model: $\Pr[Z_i = 1 \mid H_0] = \frac{1}{2}$ and $\Pr[Z_i = 1 \mid H_1] = \frac{1}{2} + \rho$ with $|\rho| \ll 1$.

The per-sentence information follows from the Bernoulli KL formula. Let $P_0 = \mathrm{Bern}(\frac{1}{2})$ and $P_1 = \mathrm{Bern}(\frac{1}{2} + \rho)$. Then $D_0^{\mathrm{S}} := D(P_1 \| P_0) = (\frac{1}{2} + \rho)\log_2(1 + 2\rho) + (\frac{1}{2} - \rho)\log_2(1 - 2\rho)$, which in the small-signal regime satisfies $D_0^{\mathrm{S}} = (1 + o(1)) \cdot 2\rho^2 / \ln 2$.

Edits occur at the token level, but semantic verification depends only on whether each evidence bit Z_i is preserved. We define the induced semantic flip rate $\varepsilon_s(\varepsilon) := \Pr[\tilde{Z}_i \neq Z_i]$, where $\tilde{Z}_i$ denotes the evidence computed from the edited sentence. If the attacker flips Z_i with probability ε_s independently, then $\Pr[\tilde{Z}_i = 1 \mid H_1] = \frac{1}{2} + (1 - 2\varepsilon_s)\rho$, while $\Pr[\tilde{Z}_i = 1 \mid H_0]$ remains $\frac{1}{2}$. Consequently, the post-edit per-sentence KL contracts quadratically:

$$D(P_{1,\varepsilon_s} \| P_{0,\varepsilon_s}) = (1 + o(1))(1 - 2\varepsilon_s)^2 \, D_0^{\mathrm{S}}. \tag{25}$$

Aggregating across T_s sentences yields $C_{\mathrm{sem}}(\varepsilon) \approx T_s(1 - 2\varepsilon_s(\varepsilon))^2 D_0^{\mathrm{S}}$. Applying Lemma E.1 gives the sufficient condition $D_0^{\mathrm{S}} \geq D_{\mathrm{req}}^{\mathrm{sem}}(\varepsilon, T_s, \beta) := \log_2(1/\beta)/[T_s(1 - 2\varepsilon_s(\varepsilon))^2]$, matching (2).

E.4. Stealth Caps from Pinsker's Inequality

Pinsker's inequality states $\mathrm{TV}(P, Q) \leq \sqrt{(\ln 2/2)\, D_{\mathrm{bits}}(Q \| P)}$. For token pooling, if an outsider pools M tokens with per-token KL drift $D_0^{(\mathrm{tok})}$, then the pooled divergence is $M D_0^{(\mathrm{tok})}$. Imposing $\mathrm{TV} \leq \tau$ yields

$$D_0^{(\mathrm{tok})} \leq D_{\mathrm{stealth}}^{\mathrm{tok}}(M, \tau) := \frac{2\tau^2}{M \ln 2}. \tag{26}$$

Similarly, pooling M_s sentences with per-sentence drift D_0^{S} and imposing $\mathrm{TV} \leq \tau_s$ yields $D_0^{\mathrm{S}} \leq D_{\mathrm{stealth}}^{\mathrm{sem}}(M_s, \tau_s) := 2\tau_s^2/(M_s \ln 2)$.

E.5. Minimal-Information Optimality

The following theorem underpins the selection rule in Section 4.

Theorem E.2 (Minimal-information principle)**.** *Fix target miss probability β and edit regime $(\varepsilon, \varepsilon_s(\varepsilon))$. Assume that (i) reliability requires $C(\varepsilon) \geq \log_2(1/\beta)$, and (ii) the detectability penalty is monotone nondecreasing in the budget D_0. Then, within any feasible family, the detectability-minimizing choice sets D_0 to the smallest value satisfying reliability, clipped by the corresponding stealth cap.*

Proof. For the token channel, Lemma E.1 combined with $C_{\mathrm{tok}}(\varepsilon) \approx T(1 - \varepsilon)^2 D_0^{(\mathrm{tok})}$ shows that meeting the power target requires $D_0^{(\mathrm{tok})} \geq D_{\mathrm{req}}^{\mathrm{tok}}$. By monotonicity of detectability, the smallest feasible budget minimizes outsider detectability. The same argument applies to the semantic channel using (25). □

E.6. Proof of the Family-Selection Rule

We now establish the selection rule from Definition 4.1. This proof optimizes over the family class considered in the main text under KL-based stealth constraints converted to total variation bounds via Pinsker's inequality. It does not cover selection over multiple drafts or joint generator-detector co-design, which lie outside our single-draft abstraction.

Feasibility sets. For the token-level and semantic families as parameterized by the per-unit KL budget D_0, feasibility requires that the required information not exceed the stealth cap. Specifically, token-level feasibility holds when $D_{\mathrm{req}}^{\mathrm{tok}}(\varepsilon, T, \beta) \leq D_{\mathrm{stealth}}^{\mathrm{tok}}(M, \tau)$. Likewise, semantic feasibility requires $D_{\mathrm{req}}^{\mathrm{sem}}(\varepsilon, T_s, \beta) \leq D_{\mathrm{stealth}}^{\mathrm{sem}}(M_s, \tau_s)$.

Distribution-preserving region. A distribution-preserving watermark achieves $\mathrm{TV} = 0$ by construction, so it dominates all probability-modifying schemes in information-theoretic stealth whenever it provides sufficient robustness. Let K denote the number of marked positions, let t denote the minimum surviving marks required, and let $X \sim \mathrm{Binomial}(K, 1 - \varepsilon)$

Table 4. Regime-conditioned within-family selection used after Definition 4.1 chooses a watermark family. For each estimated edit rate $\hat{\varepsilon}$ and watermark candidate m, the concrete scheme is the method with the highest score $S(m, \hat{\varepsilon})$.

Family	Estimated Edit Rates	Edit rate conditioned Score $S(m, \hat{\varepsilon})$ with tuple (AUC, z)	Selected Rep.	Watermarking Parameters
Token-level	$\hat{\varepsilon} \approx 0.25$ $\hat{\varepsilon}_s \approx 0.06$	**HCW: 1.137 (0.910, 3.40)** DiPMark: 1.104 (0.900, 3.90) HeavyWater: 1.072 (0.880, 4.20) SimplexWater: 1.052 (0.870, 4.50) Unigram: 0.982 (0.880, 8.80) KGW: 0.954 (0.860, 9.60)	**HCW**	δ-reweight variant (`method=delta`)
Semantic	$\hat{\varepsilon} \approx 0.42$ $\hat{\varepsilon}_s \approx 0.10$	**PMark: 1.305 (0.850, 1.20)** SemStamp: 1.220 (0.820, 1.50) SimMark: 1.170 (0.800, 1.70)	**PMark**	Rejection parameter $\rho = 0.25$
Dist.-preserving	$\hat{\varepsilon} \approx 0$ $\hat{\varepsilon}_s \approx 0$	**CGW: 1.990 (0.990, -5.80)**	**CGW**	Security parameter $\lambda = 128$, fixed secret key across runs

Note. This table explains the second stage of the Hybrid. Definition 4.1 first selects the family from the estimated edit regime. The concrete scheme is then chosen within that family using the regime-conditioned score $S(m, \hat{\varepsilon}) = \mathrm{AUC}(m, \hat{\varepsilon}) + \frac{1}{1+\max(z(m,\hat{\varepsilon}),0)}$. Greener cells indicate higher within-family scores.

count survivors. By Hoeffding's inequality, $\Pr[X < t] \le \beta$ holds whenever $(1-\varepsilon) \ge \frac{t}{K} + \sqrt{\frac{\ln(1/\beta)}{2K}}$. When this condition holds, the distribution-preserving scheme achieves target power with perfect stealth ($D_0 = 0$) and is therefore preferred within the considered family class.

Semantic and token-level branches. When distribution-preserving watermarking is infeasible, the selection rule compares the required budgets of the remaining families. If the semantic feasibility interval is nonempty, Theorem E.2 yields $D_0^{\mathrm{S}\star} = D_{\mathrm{req}}^{\mathrm{sem}}$. This branch is advantageous when ε is large but $\varepsilon_s(\varepsilon)$ is small, since $(1-2\varepsilon_s)^2$ remains close to unity even as $(1-\varepsilon)^2$ shrinks. If both families are feasible, the selection rule (5) chooses the family with the smaller required budget, thereby minimizing the Pinsker-based TV bound while meeting the robustness target. If only the token-level family is feasible, Theorem E.2 yields $D_0^{\mathrm{tok}\star} = D_{\mathrm{req}}^{\mathrm{tok}}$. □

E.7. Estimation Protocol and Robust Selection

In deployment, ε and $\varepsilon_s(\varepsilon)$ must be estimated from the anticipated editing pipeline.

Token edit-rate estimation. Let $y_{1:T}$ denote the original sequence and $\tilde{y}_{1:T'}$ the edited sequence. We define $\hat{\varepsilon} := \mathrm{EditDist}(y_{1:T}, \tilde{y}_{1:T'})/\max(T, T')$, where EditDist denotes the Levenshtein distance. Insertions and deletions are treated as equivalent to substitutions for budget purposes, which is conservative since both destroy the watermark signal. All computations use the same tokenizer as the watermarking scheme.

Semantic flip-rate estimation. We segment both original and edited text into n sentences using a deterministic rule. For each sentence i, we compute the evidence bit Z_i from the original and $\tilde{Z}_i$ from the edited sentence. The empirical flip rate is $\hat{\varepsilon}_s := n^{-1} \sum_{i=1}^{n} \mathbf{1}[Z_i \neq \tilde{Z}_i]$.

Confidence intervals. Since both estimators are averages of bounded random variables, Hoeffding's inequality yields valid bounds. For confidence level $1-\delta$, define the upper confidence bound $\varepsilon_s^{\mathrm{U}} := \hat{\varepsilon}_s + \sqrt{\frac{\ln(1/\delta)}{2n}}$, and analogously for ε^{U}.

Robustified selection. The conservative selector substitutes upper bounds into the required-budget formulas: $D_{\mathrm{req}}^{\mathrm{tok,U}} := \log_2(1/\beta)/[T(1-\varepsilon^{\mathrm{U}})^2]$ and $D_{\mathrm{req}}^{\mathrm{sem,U}} := \log_2(1/\beta)/[T_s(1-2\varepsilon_s^{\mathrm{U}})^2]$. By a union bound, if both $\varepsilon \le \varepsilon^{\mathrm{U}}$ and $\varepsilon_s \le \varepsilon_s^{\mathrm{U}}$ (which occurs with probability at least $1-2\delta$), then the selected family achieves target power $1-\beta$.

Algorithm 1 Robust Watermark Family Selection

Require: Estimated $(\hat{\varepsilon}, \hat{\varepsilon}_s)$, sample sizes (n_ε, n), confidence level $1-\delta$, target β, lengths (T, T_s), DP parameters (K, t)

Compute $\varepsilon^{\mathrm{U}} \leftarrow \hat{\varepsilon} + \sqrt{\ln(1/\delta)/(2n_\varepsilon)}$
Compute $\varepsilon_s^{\mathrm{U}} \leftarrow \hat{\varepsilon}_s + \sqrt{\ln(1/\delta)/(2n)}$
if $(1-\varepsilon^{\mathrm{U}}) \geq t/K + \sqrt{\ln(1/\beta)/(2K)}$ **then**
  return DP watermarking with $D_0 = 0$
end if
Compute $D_{\mathrm{req}}^{\mathrm{tok,U}} \leftarrow \log_2(1/\beta)/[T(1-\varepsilon^{\mathrm{U}})^2]$
Compute $D_{\mathrm{req}}^{\mathrm{sem,U}} \leftarrow \log_2(1/\beta)/[T_s(1-2\varepsilon_s^{\mathrm{U}})^2]$
if $D_{\mathrm{req}}^{\mathrm{sem,U}} < D_{\mathrm{req}}^{\mathrm{tok,U}}$ and the semantic family is feasible **then**
  return Semantic watermarking with $D_0 = D_{\mathrm{req}}^{\mathrm{sem,U}}$
else if the token-level family is feasible **then**
  return Token-level watermarking with $D_0 = D_{\mathrm{req}}^{\mathrm{tok,U}}$
else
  return Target power unattainable
end if

E.8. Sensitivity Analysis

Estimation errors in $(\varepsilon, \varepsilon_s)$ propagate to the required budgets as follows. If ε is misestimated as $\varepsilon + \Delta$, then $D_{\mathrm{req}}^{\mathrm{tok}}(\varepsilon + \Delta)/D_{\mathrm{req}}^{\mathrm{tok}}(\varepsilon) = [(1-\varepsilon)/(1-\varepsilon-\Delta)]^2$. Underestimating ε causes the required budget to be underestimated, potentially leading to insufficient power. An analogous relation holds for ε_s: $D_{\mathrm{req}}^{\mathrm{sem}}(\varepsilon_s+\Delta_s)/D_{\mathrm{req}}^{\mathrm{sem}}(\varepsilon_s) = [(1-2\varepsilon_s)/(1-2\varepsilon_s-2\Delta_s)]^2$.

With Hoeffding-based confidence intervals at level $1-\delta$ and sample size n, we have $\Delta_\varepsilon, \Delta_s = O(\sqrt{\ln(1/\delta)/n})$, so the regret factor from using upper bounds converges to 1 as $n \to \infty$. Here, regret is measured with respect to the paper's selection objective, namely the minimum required KL budget among feasible families in the considered class, rather than regret over all conceivable watermarking strategies. This analysis demonstrates that the robust selector's conservatism is bounded and diminishes with additional calibration data.

F. Watermarks Beyond Sampling

Our theorems are proved for sampling-based watermarks; however, detectability depends only on the total variation between the unwatermarked distribution P and the watermarked distribution Q, regardless of implementation. Weight-embedded or structural watermarks fit the same two-hypothesis framework whenever they induce $Q_{\theta+\xi} \neq P_\theta$. The relevant parameter is the per-token KL signal $D_0 \equiv \mathbb{E}[D(q_t \| p_t)]$, which for structural schemes represents the average effect of parameter perturbation on next-token distributions. Any mechanism producing a nonzero D_0 falls under Theorem 3.2, while Theorem 3.4 and the attenuation law $D_\varepsilon \approx (1-\varepsilon)^2 D_0$ apply when adversaries act on outputs. Our theory does not address robustness to parameter-space transformations such as fine-tuning, distillation, or pruning.

Recent work (Gu et al., 2024) embeds watermarks directly into model parameters. It demonstrates that models can be trained to produce text satisfying watermark predicates, effectively learning to generate watermarked outputs without explicit sampling modifications. If the modified model produces conditionals $q_t \neq p_t$, our results apply unchanged; if $q_t \approx p_t$, then $D_0 \approx 0$ and no text-level signal exists for black-box detection.

Generator-side selection (WaterMax-style) in our framework. Methods that generate multiple candidates and select the best according to a keyed score, such as WaterMax (Giboulot & Furon, 2024), define an *implicit* output distribution Q that typically differs from the baseline P, thereby inducing a nonzero KL budget D_0. Such methods can improve the *robustness–quality* tradeoff by expending additional compute (generating more candidates) rather than increasing token-level distortion; however, the resulting detectability is still governed by the induced separation between Q and P. Our empirical utility and overhead results (Appendix G.7) capture this compute-versus-detectability trade-off, showing that semantic-rejection-based methods incur 2.4-2.8$\times$ overhead while achieving comparable or superior robustness to token-level schemes.

Table 5. Low-entropy code-domain stress test on MBPP using a pretrained CodeLlama checkpoint. Results are reported on 50 MBPP tasks with 10 samples per task. Detection statistic denotes the mean detector output (z-score when available).

Method	pass@1	pass@10	Parse Rate	Detection Rate	Mean Detection Statistic	Interpretation
Vanilla	0.4	0.420	0.328	0.000	0.00	Baseline low-entropy reference point.
KGW	0.28	0.340	0.256	0.804	z = 7.42	Clear utility drop with strong keyless detectability.
Unigram	0.26	0.320	0.298	0.816	z = 8.30	Largest utility drop among measured methods, with strong detectability.
Hybrid	0.32	0.420	0.290	0.454	z = 2.29	Best utility preservation among detectable methods; moderate detectability.

Note. The code setting exhibits substantially lower next-token uncertainty than the open-ended LFQA setting, evaluated as mean next-token entropy of 0.93 bits on MBPP versus 2.24 bits on LFQA. Accordingly, utility degradation in MBPP is better reflected by diversity-sensitive metrics such as pass@k than by text similarity alone.

Spectral and correlation-based watermarks. Some recent methods embed signals in higher-order statistics (e.g., correlations across positions) rather than per-token bias. These methods still fit the two-hypothesis view via the sequence distributions $P(\mathbf{y})$ and $Q(\mathbf{y})$, but their detectability may not be well summarized by a per-token KL budget D_0 alone. Extending our analysis to block-level KL (or mixing-time-controlled dependence) is an interesting direction; empirically, our keyless detector suite includes correlation-sensitive tests that partially probe this regime.

Key-averaged distributions. For schemes with multiple keys $k \in \mathcal{K}$, an adversary observing outputs under different keys sees the key-averaged distribution $\overline{Q} = \mathbb{E}_k[Q_k]$, and Theorem 3.2 applies with this substitution. Distribution-preserving schemes yield $\overline{Q} = P^s$ and zero TV, while probability-modifying schemes generally maintain separation from P^s.

G. Additional Experimental Results

This appendix extends our empirical evaluation with additional models, attack types, and detailed analyses that support the theoretical framework presented in the main text. We organize the material as follows: we first define evaluation metrics and their interpretation in the watermarking context; Table 6 then presents results on Mistral-7B; subsequent subsections detail the attack catalog, estimation protocols, hybrid selector sensitivity, adaptive attack evaluation, utility metrics, and validation of theoretical assumptions.

G.1. Evaluation Metrics and Their Interpretation

Watermarking schemes face a fundamental tension between multiple desiderata: the watermark should be *robust* (detectable by the authorized verifier even after text modifications), yet *undetectable* by unauthorized third parties, while preserving the *utility* (quality) of generated text. We evaluate schemes along these three axes using standard metrics from the detection theory and natural language generation literature.

Robustness metrics. Robustness measures how reliably an authorized verifier with the secret key can identify watermarked text, particularly after the text has been edited or paraphrased. We report two complementary metrics. *AUROC* (Area Under the Receiver Operating Characteristic curve) measures the probability that a randomly chosen watermarked sample receives a higher verification score than a randomly chosen unwatermarked sample. An AUROC of 1.0 indicates perfect separation between watermarked and unwatermarked distributions, while an AUROC of 0.5 indicates performance equivalent to random guessing. *TPR at 1% FPR* (True Positive Rate at 1% False Positive Rate) measures the fraction of watermarked texts correctly identified when the verification threshold is set to produce at most 1% false accusations of unwatermarked text. This metric is particularly relevant for deployment scenarios where falsely accusing human-written text of being AI-generated carries significant consequences.

Detectability metrics. Detectability measures how easily an unauthorized third party without the secret key can identify watermarked text using only statistical analysis. Low detectability is desirable in applications requiring plausible deniability. We focus on the *z-score* detector of (Liu et al., 2025), which measures the deviation of observed token statistics from their expected values under the null hypothesis of no watermarking. Under the null hypothesis, z-scores follow approximately

Table 6. Robustness and detectability on Mistral-7B across attack conditions. For each condition, we report AUROC (AUC), TPR at 1% FPR, and keyless z-score. Superscripts denote families as in Table 3.

	No attack			DIPPER ($\hat{\varepsilon}\approx0.25$)			OPT-2.7B ($\hat{\varepsilon}\approx0.15$)			WM-removal ($\hat{\varepsilon}\approx0.15$)			Synonym ($\hat{\varepsilon}\approx0.15$)			Back-trans. ($\hat{\varepsilon}\approx0.42$)		
Method	AUC	TPR	z	AUC	TPR	z	AUC	TPR	z	AUC	TPR	z	AUC	TPR	z	AUC	TPR	z
KGW^{B}	.99	1.00	27.8	.85	.62	9.0	.76	.57	8.2	.76	.56	7.9	.76	.57	8.0	.56	.50	−1.4
Unigram^{B}	.99	1.00	10.5	.86	.64	8.2	.77	.59	7.7	.77	.58	7.4	.77	.59	7.5	.55	.49	−1.2
DiPMark^{F}	.99	1.00	39.5	.89	.79	4.1	.89	.84	3.8	.88	.83	3.5	.89	.84	3.6	.58	.52	−1.0
HCW^{F}	.99	1.00	98.7	.90	.80	3.5	.91	.87	3.2	.90	.86	3.1	.91	.87	3.1	.57	.51	−0.8
HeavyWater^{F}	.99	1.00	36.2	.86	.73	4.4	.87	.79	4.0	.86	.78	3.7	.87	.79	3.8	.54	.48	−1.2
SimplexWater^{F}	.99	1.00	33.8	.85	.71	4.7	.86	.77	4.2	.85	.76	3.9	.86	.77	4.0	.53	.47	−1.3
SemStamp^{S}	.98	.98	7.8	.91	.86	2.6	.93	.90	2.2	.92	.89	2.0	.93	.90	2.1	.80	.73	1.3
PMark^{S}	.99	.99	6.5	.93	.87	2.3	.94	.91	1.9	.93	.90	1.7	.94	.91	1.8	.83	.77	1.0
SimMark^{S}	.98	.97	8.4	.90	.83	2.9	.92	.88	2.5	.91	.87	2.3	.92	.88	2.4	.78	.70	1.5
CGW^{D}	.99	1.00	−12.5	.50	.14	−8.9	.50	.29	−9.7	.50	.29	−9.8	.50	.29	−9.6	.50	.19	−7.0
GaussMark^{W}	1.00	1.00	11.8	.80	.68	8.2	.83	.73	7.5	.81	.71	7.1	.82	.72	7.2	.60	.53	1.9
Hybrid*	.99	1.00	−11.0	.93	.88	4.0	.95	.92	3.6	.94	.91	3.8	.94	.92	3.9	.84	.77	1.6

a standard normal distribution with mean 0 and standard deviation 1. A z-score of 3.0 corresponds to a p-value of approximately 0.001, indicating strong statistical evidence of watermarking. For distribution-preserving schemes like CGW, z-scores should remain close to the null distribution even after text is watermarked. As an alternative detector-side approach, we note the Tr-GoF detector of (Li et al., 2025), which targets edited text by reframing detection as a sparse-mixture problem and using a truncated goodness-of-fit statistic. This detector is best viewed as a verifier-side improvement for distribution-preserving and unbiased watermarks under human edits; we discuss it here to contextualize detector choice while keeping the main text focused on the z-score scheme (Liu et al., 2025).

Utility metrics. Utility measures how much the watermarking process degrades the quality of the generated text. *MAUVE* (Pillutla et al., 2021) measures the distributional similarity between watermarked and unwatermarked text collections using neural features. *BERTScore* (Zhang et al., 2019) measures semantic similarity between watermarked and unwatermarked outputs for the same prompts using contextual embeddings. *LLM-as-judge* uses GPT-4 to rate the helpfulness and coherence of watermarked outputs on a 1 to 5 scale, following established evaluation protocols (Zheng et al., 2023).

Edit rate metrics. To characterize attack strength and connect experimental results to our theoretical analysis, we measure two edit rates. The *token edit rate* $\hat{\varepsilon}$ measures the fraction of tokens modified by an attack, computed as the normalized Levenshtein distance between original and edited token sequences. The *semantic flip rate* $\hat{\varepsilon}_s$ measures the fraction of sentences whose watermark evidence changes after editing. Critically, the semantic flip rate is often much lower than the token edit rate for meaning-preserving attacks.

G.2. Results on Mistral-7B

Table 6 presents results on Mistral-7B using the same experimental protocol applied to Llama-2-7B in the main text. The robustness and detectability tradeoff on Mistral-7B closely mirrors the patterns observed on Llama-2-7B (Table 3), indicating that this tradeoff is governed by the watermarking family rather than the underlying language model. In the reference condition, all methods achieve near-perfect robustness while exhibiting consistent detectability ordering: CGW sits near the low-detectability corner, semantic watermarks achieve similarly low detectability while maintaining strong robustness, and biased schemes are easily flagged statistically. Under paraphrasing attacks, the relative positioning of schemes remains consistent across both models, with semantic watermarks demonstrating superior robustness compared to token-level methods at equivalent detectability levels.

G.3. Attack Catalog and Edit Regime Measurements

Table 7 catalogs all attacks evaluated in our experiments, reporting the realized token edit rate $\hat{\varepsilon}$ and the measured semantic flip rate $\hat{\varepsilon}_s$ for semantic watermarking schemes. These measurements operationalize the key parameters in Theorem 3.4.

The table reveals several important patterns. First, oblivious paraphrasers (DIPPER, OPT-2.7B, synonym substitution) induce semantic flip rates substantially lower than their token edit rates ($\hat{\varepsilon}_s \ll \hat{\varepsilon}$), confirming the regime where semantic watermarks outperform token-level schemes. Second, back-translation produces high edit rates but preserves semantic

Table 7. Attack catalog with measured edit regimes. All values are mean ± 95% confidence interval over 500 prompts.

Attack	$\hat{\varepsilon}$	$\hat{\varepsilon}_s^{\text{PMark}}$	**Notes**
No attack	0.00	0.00	Baseline
DIPPER	0.25±0.02	0.06±0.01	Calibrated
OPT-2.7B	0.15±0.03	0.04±0.01	Prompted
Synonym sub.	0.15±0.02	0.02±0.01	Lexical only
WM-removal	0.15±0.02	0.05±0.01	Oblivious
Back-trans.	0.42±0.08	0.10±0.02	Semantic
Key-aware	0.15±0.01	0.09±0.02	Targeted
Detector-guided	0.18±0.03	0.05±0.01	$K = 10$

Table 8. Hybrid selection regret under parameter misestimation.

Selector	**Avg loss**	**Wrong family**	**Worst case**
Oracle	0.000	0%	0.000
Plug-in	0.008±0.003	4.2%	0.031
Conservative	0.011±0.004	2.1%	0.018

structure, explaining the performance gap between watermarking families in Table 3. Third, adaptive attacks (key-aware and detector-guided) achieve higher semantic disruption per token edit, operating closer to the regime boundary where token-level and semantic schemes have comparable robustness.

G.4. Edit Rate Estimation Protocol

Accurate estimation of edit parameters is essential for the hybrid selector to operate correctly. For the token edit rate, we compute $\hat{\varepsilon} = \text{EditDistance}(\mathbf{y}, \tilde{\mathbf{y}}) / \max(T, T')$ where $\mathbf{y}$ and $\tilde{\mathbf{y}}$ are the original and edited token sequences with lengths T and T', and EditDistance is the Levenshtein distance. For the semantic flip rate, we segment text into sentences using spaCy and compute the watermark evidence bit $Z_t \in \{0, 1\}$ for each sentence before and after editing. The flip rate is $\hat{\varepsilon}_s = T_s^{-1} \sum_{t=1}^{T_s} \mathbf{1}[\widetilde{Z}_t \neq Z_t]$ where T_s is the number of aligned sentence pairs. Sentences without clear alignment due to splitting or merging are treated as flipped.

The robust hybrid selector uses upper confidence bounds rather than point estimates. Given the edited sample pair, the selector estimates both $\hat{\varepsilon}$ and $\hat{\varepsilon}_s$ with their confidence intervals, computes the required detectability for both token-level and semantic families using Theorem 3.4 evaluated at the upper bounds, and selects the family achieving the lower detectability requirement.

G.5. Hybrid Selector Sensitivity Analysis

We quantify sensitivity to estimation error by comparing three selection strategies: an **oracle** using true parameter values, a **plug-in selector** using point estimates, and a **conservative selector** using upper confidence bounds. Table 8 reports the regret (AUROC loss relative to oracle) under systematic estimation perturbations of ± 0.05 and ± 0.10 in $\hat{\varepsilon}_s$.

The plug-in selector achieves low average regret but occasionally selects the wrong family near regime boundaries, leading to worst-case AUROC losses up to 0.031. The conservative selector trades slightly higher average regret for substantially reduced worst-case regret (0.018 versus 0.031), making it preferable for risk-averse deployments.

G.6. Adaptive Attack Evaluation

We evaluate two adaptive adversaries substantially stronger than oblivious paraphrasers. The **key-aware attack** assumes access to the secret key and greedily edits tokens that contribute most to the verification score, subject to maintaining semantic similarity (BERTScore (Zhang et al., 2019) ≥ 0.85). The **detector-guided attack** models a public adversary without key access: it generates $K = 10$ paraphrase candidates and selects the one minimizing a keyless detectability score subject to the same similarity constraint.

Adaptive attacks substantially degrade token-level schemes, with AUROC drops of 0.16 to 0.18 for key-aware attacks. Semantic schemes are more resilient, with drops of only 0.08 to 0.10. The hybrid selector adapts appropriately, achieving

Table 9. Robustness (AUROC) under oblivious versus adaptive attacks at $\hat{\varepsilon} \approx 0.15$.

Method	Oblivious	Key-aware	Det.-guided
KGW^{B}	.78	.61	.69
DiPMark^{F}	.91	.74	.82
HeavyWater^{F}	.89	.71	.79
SemStamp^{S}	.94	.85	.91
PMark^{S}	.95	.87	.93
SimMark^{S}	.93	.83	.90
Hybrid*	.96	.88	.94

Table 10. Utility and compute overhead. Higher MAUVE, BERTScore, and LLM-judge indicate better quality.

Method	MAUVE	BERTScore	LLM	Cost
Unwatermarked	1.00	1.00	4.21	1.0×
KGW	0.91	0.96	4.08	1.0×
DiPMark	0.95	0.98	4.15	1.1×
HeavyWater	0.97	0.99	4.18	1.2×
SimplexWater	0.97	0.99	4.17	1.2×
SemStamp	0.94	0.97	4.14	2.8×
PMark	0.95	0.98	4.17	2.4×
SimMark	0.94	0.97	4.15	2.6×
CGW	0.99	0.99	4.20	1.5×
Hybrid	0.95	0.98	4.16	1.8×

robustness comparable to the best single-family scheme under each attack type.

G.7. Utility and Compute Overhead

We report utility metrics in Table 10. Distribution-preserving CGW achieves near-perfect utility (MAUVE 0.99) but sacrifices robustness. Biased schemes show modest utility degradation (MAUVE 0.91) due to distribution shift. Semantic schemes incur compute overhead (2.4 to 2.8×) due to rejection sampling. The hybrid achieves utility comparable to bias-free methods with moderate overhead (1.8×), invoking semantic methods only when necessary.

G.8. Keyless Detector Baseline

For interpretability of keyless detection scores, we report baseline statistics on unwatermarked Llama-2-7B outputs over 500 prompts. The z-score has mean -0.3 with standard deviation 2.1, centered near zero as expected under the null hypothesis of no watermarking. CGW's z-scores (-5.8 to -10 in Table 3) fall within 2 to 3 standard deviations of this baseline, consistent with distribution preservation.

G.9. Validation of Theoretical Assumptions

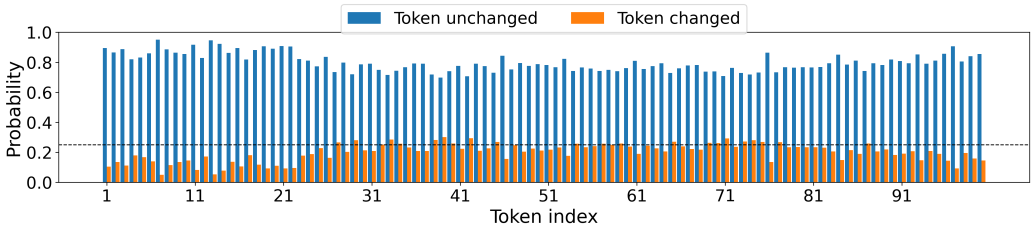
Our theoretical analysis relies on two key assumptions: quadratic approximations for KL divergence in the small-bias regime, and approximate independence of edits across positions. We validate both assumptions empirically to establish the practical applicability of our theoretical bounds.

KL approximation accuracy. For biased sampling schemes with green-list bias parameter δ, the per-token KL divergence admits the quadratic approximation $D(q_t \| p_t) \approx \delta^2 g_t(1 - g_t)/(2 \ln 2)$, where g_t denotes the green-list probability mass at position t. This approximation arises from a second-order Taylor expansion of the KL divergence around the unbiased distribution and holds when the bias-induced probability shift remains small relative to the baseline token probabilities. We validated this approximation at our experimental hyperparameters by comparing exact KL values, computed via numerical integration over the vocabulary, against the quadratic formula across all token positions in a representative sample of 500 generated sequences.

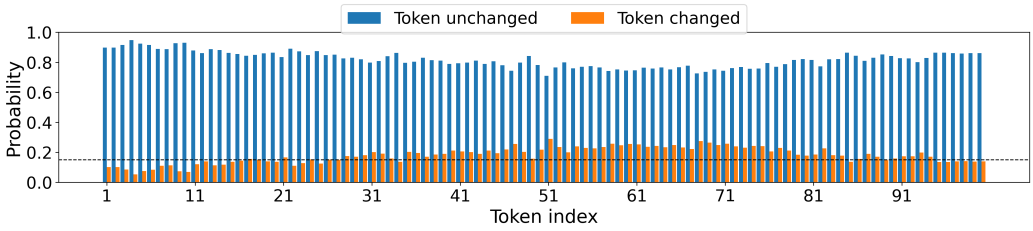
For KGW with $\delta = 2.0$, the median relative error between exact and approximate KL is 8.2%, with 90% of positions exhibiting errors below 15%. For DiPMark with $\delta = 1.0$, the smaller bias yields improved approximation accuracy,

Table 11. Robustness under independent versus correlated edits at $\hat{\varepsilon} \approx 0.20$.

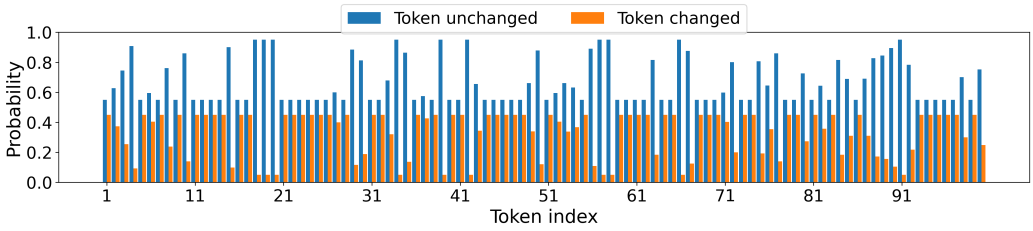
Method	Independent	Correlated	ΔAUROC
KGW	.86	.83	−.03
DiPMark	.90	.87	−.03
PMark	.94	.92	−.02
Hybrid	.94	.92	−.02



(a) DIPPER ($\varepsilon \approx 0.25$)



(b) Synonym substitution ($\varepsilon \approx 0.15$)



(c) Back-translation (unconstrained)

Figure 3. Per-token modification probabilities. For ten 100-token outputs, we apply each attack 10 times and estimate the probability of modification (orange) or preservation (blue) at each position. Dashed lines show global edit rate where applicable.

reducing the median error to 2.8%. PMark with rejection parameter $\rho = 0.25$ achieves a median error of 3.2%, reflecting its distribution-modifying mechanism that operates through selective rejection rather than explicit biasing. The quadratic approximation achieves median relative errors below 10% for all schemes at their operational hyperparameters, confirming sufficient accuracy for our theoretical analysis. We note that the approximation degrades gracefully as δ increases: at $\delta = 4.0$, the median error rises to approximately 18%, which remains acceptable for order-of-magnitude predictions but suggests that extremely aggressive biasing may require higher-order corrections.

Edit channel independence. The robustness bounds in Theorem 3.4 model post-editing observations as passing through a binary symmetric channel with independent flips at rate ε. Real attacks may induce correlated edits, particularly when operating on contiguous spans. To assess the impact of such correlations, we stress-test the independence assumption using a **span rewriting** attack that systematically rewrites contiguous spans of 5 to 10 tokens, inducing strong local correlation within each span while maintaining approximate independence across spans.

We compare detection performance under span rewriting against a baseline of independent random substitutions calibrated to the same aggregate edit rate. Correlated edits produce modest additional degradation of 0.02 to 0.03 AUROC compared to independent edits at matched rates. This degradation is consistent with the theoretical intuition that correlation reduces the effective sample size: a span of k correlated edits contributes less information than k independent edits but more than a single edit. Crucially, the threshold predictions from Theorem 3.4 remain valid under span rewriting, as the dominant factor

Table 12. GaussMark detection power at $\alpha = 0.05$ as a function of σ (held-out calibration). The detection statistic ψ measures the normalized correlation between output logits and the secret perturbation direction.

σ	TPR	FPR	Avg. ψ
0.00	0.00	0.00	0.002
0.01	0.60	0.00	0.079
0.03	0.80	0.00	0.118
0.05	1.00	0.00	0.168
0.07	1.00	0.00	0.204
0.10	1.00	0.20	0.254

governing detection failure is the aggregate edit rate rather than the fine-grained correlation structure. This robustness to correlation violations stems from the concentration behavior of the detection statistic, which depends primarily on the total number of surviving watermark signals rather than their spatial arrangement.

Figure 3 visualizes per-position edit probabilities under different attack strategies, providing direct evidence for the independence assumption. DIPPER produces a nearly uniform edit profile across positions, with position-wise edit rates exhibiting low variance (coefficient of variation below 0.15), strongly supporting our independence assumption. Synonym substitution shows greater positional variability, concentrating edits on content words while preserving function words, but maintains approximate stationarity across the sequence. Back-translation produces localized spikes corresponding to phrases that the round-trip translation restructures, violating local independence while preserving global stationarity.

For paraphrasers that maintain approximately constant edit rates across positions, our first-order independent model provides a sound approximation. The key insight is that detection statistics aggregate information across many positions, and by the law of large numbers, local correlations average out when the correlation length is short relative to the sequence length. Back-translation's unconstrained rewriting produces substantially higher aggregate edit rates (often exceeding 60%), placing it in the high-noise regime where Theorem 3.4 predicts detection failure regardless of correlation structure. The experimental results in Table 3 confirm this prediction: all watermarking schemes fail to detect back-translated text, consistent with the theoretical bound rather than with correlation-induced degradation.

G.10. GaussMark Hyperparameter Calibration

Following GaussMark (Block et al., 2025), we tune the Gaussian perturbation scale σ on a held-out split by sweeping σ and measuring detection power at fixed significance $\alpha = 0.05$. Because GaussMark is a training-time method that requires gradient access for verification, it is not directly comparable to inference-time watermarks in our hierarchy; we include this calibration for completeness. At $\sigma = 0.05$, GaussMark achieves perfect detection (TPR = 1.00) at the target FPR. Higher values of σ (Table 12) increase the detection statistic ψ but begin to incur false positives, reflecting the tradeoff between embedding strength and utility degradation that parallels the robustness-detectability tradeoff in inference-time watermarks.