SIMPLEMARK: FAST MULTI-BIT TEXT WATERMARKING UNDER f-DIVERGENCE CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a framework for multi-bit text watermarking with security defined directly through f-divergence from the base language model distribution. Unlike prior approaches that focus on average-key distortion-freeness or a particular statistical distance, our formulation supports general f-divergences, including total variation and KL divergence, and enforces the guarantee for each realized key and embedded message. We develop a coding-based watermarking scheme that optimally biases next-token distributions subject to a prescribed divergence budget, and characterize the resulting tradeoff between embedding rate, decoding reliability, and statistical security. Experimentally, we compare our method against prior multi-bit watermarking schemes across modern language models and payload regimes. Our approach achieves substantially lower watermark detectability while maintaining competitive message-recovery performance and generation quality. Our results provide a unified view of secure multi-bit watermarking and recover several commonly used security notions as special cases.

1 INTRODUCTION

Large language models (LLMs) can now generate fluent, high-quality text at scale.This creates new challenges for provenance, accountability, and policy enforcement (Longpre et al., 2024). Text watermarking offers a promising response: a model provider can embed a hidden signal into generated text so that the output can later be detected or decoded (Kirchenbauer et al., 2023). While early watermarking schemes primarily focus on *zero-bit* detection (to ascertain the presence of a watermark) (Christ et al., 2023; Kuditipudi et al., 2024; Kirchenbauer et al., 2023), many applications require *multi-bit* watermarking, where the generated text carries a message such as a user identifier, model identifier, timestamp, or policy tag (Cox et al., 2007).

Multi-bit watermarking introduces a fundamental tension: increasing the payload generally requires a stronger perturbation of the model's next-token distribution, making the watermark both easier to recover and easier to detect by an adversary (Jiang et al., 2025). Conversely, imposing stronger distributional constraints on the watermarked model, relative to the unwatermarked model, makes the watermark less detectable but limits the amount of information that can be reliably embedded. This raises a basic question: *what are the capacity and detectability of a multi-bit text watermark while being as close as possible to the unwatermarked distribution?* We study this question by comparing how far from the unwatermarked distributions different watermarks perturb down to individual token distributions and fundamentally how much information one can embed. To measure distance from the base model, we adopt per-key rather than average-key notions (Gilani et al., 2026). This is because in adversarial settings (Chang et al., 2025; Zhang et al., 2024; Adelsbach et al., 2003), security should hold for the actual distribution produced in a realized execution, not after averaging over secret keys (Katz & Lindell, 2014). An average-key guarantee can mask large key-dependent deviations that give a distinguisher substantial advantage, motivating a stronger per-key notion of statistical indistinguishability (Cox et al., 2007).

Motivated by this observation, we introduce SIMPLEMARK, a multi-bit text watermarking scheme that enforces an f-divergence ε-budget at each generation step. SIMPLEMARK uses low density parity check (LDPC) codes with a geometry-aware (simplex) optimal transport embedding rule and maximum-likelihood decoding. The simplex geometry allows the reduction of the per-token embedding step to a two-scalar tilt computable in $O(1)$, while the LDPC coding allows for improved,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

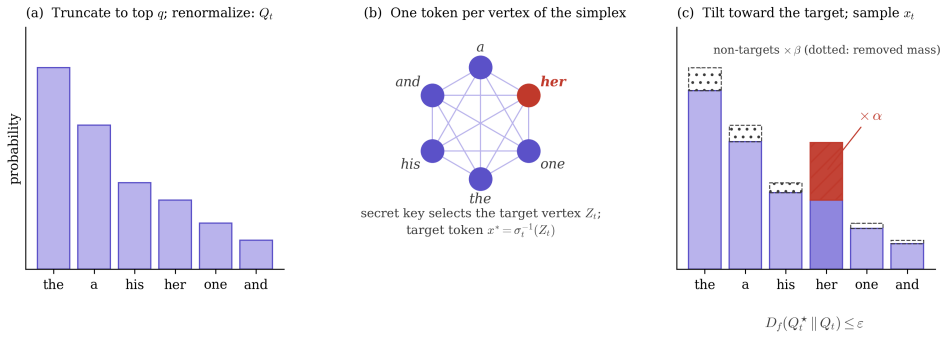

Figure 1: Watermark encoding and embedding process for SIMPLEMARK: the watermarker starts by truncating the language model support according to the top-q, and places all the tokens on a $q-$ary simplex. Using the message M and the generator matrix G to create a codeword C, and computes the target vertex on the simplex, depicted in red in the diagram. The watermarker biases the watermarked distribution towards this token by α, and reduces the probability of sampling other tokens by β while remaining in the ε parameter f-divergence budget, $D_f(Q_t^*\|Q_t) \leq \varepsilon$, and samples from the watermarked distribution Q_t^*. The message can then be decoded with one of the decoders. A detailed procedure is in Section 3.1.

coding-theoretic reliability in recovery. Figure 1 graphically illustrates SIMPLEMARK. In Section 3.1, we discuss the components of SIMPLEMARK in more detail.

Our contributions: First, we formalize multi-bit text watermarking as communication through a constrained watermark channel and identify the mutual information governing reliable message recovery. Second, we show a channel-collapse result: exact per-key distortion-freeness forces zero embedding capacity (i.e., no scheme can embed even a single bit) so distortion is not a defect of existing multi-bit watermarks but the price of embedding, and the right goal is to control it explicitly, through f-divergence constraints yielding a tunable capacity–detectability tradeoff. Third, we demonstrate that average-key distortion-freeness and other such metrics do not imply per-key undetectability and can lead to highly detectable realized-key behavior. Finally, we evaluate SIMPLEMARK on modern language models showing it achieves competitive or state-of-the-art performance on reasoning and advanced tasks, recovery, perplexity, and more while substantially reducing the distributional deviation between the watermarked model and the base model.

1.1 RELATED WORK

Multi-bit Watermarks Recently, a variety of multi-bit watermarks have been developed for language models (Yoo et al., 2024; Jiang et al., 2025; Qu et al., 2025; Xu et al., 2025; Jiang et al., 2026; Cui et al., 2026). Two prominent ones are BiMark (Feng et al., 2025) and ArcMark (Gilani et al., 2026). A pivotal advancement is the development of provably robust frameworks; Qu et al. (2025) use error-correction codes (ECC) to provide formal guarantees within specific edit-distance bounds, whereas (Cohen et al., 2025) establishes resistance against adaptive prompting. Recent work such as (Xu et al., 2026) further optimizes this tradeoff by employing block-wise encoding to ensure high recovery rates even in short-form textual outputs, while (Guo et al., 2026) incorporates feedback coding. To the best of our knowledge, these previous works do not provide an explicit per-key mechanism for directly tuning a prescribed divergence budget (e.g., for total variation or KL) at each generation step.

Watermarking Capacity Our work connects to the information-theoretic literature of watermarking, data hiding, and steganography, where watermarking is modeled as communication over a constrained channel: the encoder embeds a message into a host signal subject to a distortion or detectability constraint, and the decoder attempts to recover it (Cachin, 2004; Cox et al., 2007). Early work characterized achievable rates for information hiding under distortion and attack constraints (Moulin & O'Sullivan, 2003; Moulin, 2003; 2005; Moulin & Koetter, 2005), and coding-based

Watermark	Per-token watermark-specific overhead	Decoding cost
ArcMark	$O(I\,\lvert\mathcal{V}\rvert\,r)$	$O\big(n(\lvert\mathcal{V}\rvert + 2^b)\big)$
BiMark	$O(d\,\lvert\mathcal{V}\rvert)$	$O(nd + b)$
StealthInk	$O(\lvert\mathcal{V}\rvert)$	$O(n\lvert\mathcal{V}\rvert + (b/c)2^c)$
MC^{2}Mark	$O\Big(m\lvert\mathcal{V}\rvert + m\binom{n'}{n'/2}n'\Big)$	$O(n\,m\,n' + b)$
QuantileMark	$O(\lvert\mathcal{V}\rvert \log \lvert\mathcal{V}\rvert + M)$	$O(n(\lvert\mathcal{V}\rvert \log \lvert\mathcal{V}\rvert + M) + (b/m')M)^\dagger$
MirrorMark	$O(L\,\lvert\mathcal{V}\rvert)$	$O(n(b/c + 2^c L))$
SIMPLEMARK	$O_q(1)$	$O\big(n\lvert\mathcal{V}\rvert + 2^b n\big)^\dagger$

Table 1: Complexity comparison between SIMPLEMARK and other known multi-bit watermark schemes. †Note the decoder requires a model replay. We assume $q = 53$ as a small constant. With respect to the complexities, the symbols can be defined as $|\mathcal{V}|$: vocabulary; n: text length; b: payload bits. Method-specific symbols: r: side-information resolution and I: Sinkhorn iteration count for ArcMark (Gilani et al., 2026); d: reweighting layers for BiMark (Feng et al., 2025) (d=10); m: channel layers and n': segment length in bits for MC^{2}Mark (Cui et al., 2026) (m=10; n' unspecified in their experiments); c: bits per message chunk for StealthInk (Jiang et al., 2025) and MirrorMark (Jiang et al., 2026) (c=1 and c=2 respectively; both papers denote this m); L: tournament layers for MirrorMark (L=30). M: quantile bins per step for QuantileMark (Zhu et al., 2026) ($M = 2^{m'}$ with m' bits per message symbol).

schemes such as quantization-index modulation showed how structured codes can approach these limits (Chen & Wornell, 2001).

In the language model setting, the host distribution is the base model's next-token distribution, so the relevant constraint is distributional rather than perceptual. Recent work has begun to analyze LLM watermarking through this lens (Gilani et al., 2026; He et al., 2025). We further extend this capacity viewpoint of multi-bit text watermarking by comparing average-key distortion-freeness, uniform per-key total-variation closeness, and perfect distortion-freeness. Our results show that stronger per-key distortion-freeness constraints reduce the mutual information of the induced watermark channel, and exact per-key distortion-freeness forces zero embedding capacity.

2 PRELIMINARIES AND NOTATION

First, we introduce some useful notation and background, many of which can be found in standard textbooks (Ross, 1998; Gut, 2013; Mitzenmacher & Upfal, 2005; Keener, 2010).

f-divergences. Let $f : (0,\infty) \to \mathbb{R}$ be convex with $f(1) = 0$. For distributions P, Q on $\mathcal{X}$ with $P \ll Q$, the f-divergence of P from Q is $D_f(P\|Q) := \sum_{x\in\mathcal{X}} Q(x)\, f\big(\frac{P(x)}{Q(x)}\big)$. Every f-divergence satisfies the data-processing inequality: $D_f(\kappa(P)\|\kappa(Q)) \le D_f(P\|Q)$ for any (possibly randomized) map κ (Polyanskiy & Wu, 2025). We next discuss the two f-divergences this work focuses on.

Total variation distance. For probability measures P, Q on measurable space $(\mathcal{X}, \mathcal{F})$, we define $d_{\mathrm{TV}}(P, Q) := \sup_{A\in\mathcal{F}} |P(A) - Q(A)| = \frac{1}{2}\sum_{x\in\mathcal{X}} |P(x) - Q(x)|$. Equivalently, total variation distance characterizes the maximum distinguishing advantage of any adversarial attacker trying to decide whether a sample was drawn from P or Q (Katz & Lindell, 2014). In particular, for any possibly randomized distinguisher $\mathcal{D} : \mathcal{X} \to \{0, 1\}$,

$$\sup_{\mathcal{D}} \left| \Pr_{X\sim P}[\mathcal{D}(X) = 1] - \Pr_{X\sim Q}[\mathcal{D}(X) = 1] \right| = d_{\mathrm{TV}}(P, Q).$$

Relating to this work, we want the TV distance of the watermarked and unwatermarked distributions to be as small as possible, so an adversary has the smallest possible advantage in determining whether outputted text is watermarked or not.

Kullback–Leibler divergence. For distributions P, Q on a finite alphabet $\mathcal{X}$ with $P \ll Q$, the Kullback–Leibler divergence of P from Q is $D_{KL}(P\|Q) := \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$. In our setting, $KL(P\|Q)$ is the expected excess log-loss incurred when text generated by the watermarked distribution P is scored by the base model Q, and thus relates directly to perplexity: writing $H(P)$ for the entropy of P, $\mathrm{PPL} = \exp\big(\mathbb{E}_{x\sim P}[-\log Q(x)]\big) = e^{H(P)}\, e^{KL(P\|Q)}$. Here, $e^{KL(P\|Q)}$ is related to the multiplicative perplexity inflation introduced by watermarking.

Watermarked distribution. Fix $\varepsilon \geq 0$. A watermarked distribution $Q_t^\star = P_{X_t|M=m,V_t=v}$ is said to be (uniformly per-key) *ε-distortion-free* if for every time t, every message $m \in \mathcal{M}$, and every realized key value v, $D_f\big(P_{X_t|M=m,V_t=v} \,\|\, Q_t\big) \leq \varepsilon$. Equivalently, when the encoder depends on (m, v) only through the state $Z_t = z_t(m, v)$, the distribution is (uniformly per-key) *ε-distortion-free* iff $D_f\big(P_{X_t|Z_t=z} \,\|\, Q_t\big) \leq \varepsilon$ for every realizable $z \in \mathcal{Z},\ \forall t$.

For our watermark, we can set ε to be a constant or to be cryptographically negligible (depending on the target security application). When $\varepsilon = 0$, we obtain *perfect* distortion-freeness.

3 Main Results

This section contains our main theoretical results. We begin by describing SimpleMark's encoding, embedding, and decoding algorithms.

3.1 Algorithms

We describe SimpleMark being used in practice. Its components include:

- **Encoder:** The message is coded with an LDPC code and then combined with the shared secret to compute the target vertex.
- **Embedding:** at each time t, the watermarked distribution $Q_t^\star$ is computed in closed form while adhering to the divergence budget. The watermarked token X_t is sampled.
- **Decoding:** a decoder uses a likelihood decoder to compute $\hat{M}$. Alternatively, we provide a plurality decoder where only the received symbols are required.

Encoding and embedding are given in algorithm form in Appendix A.

Message Encoding The watermarker begins with a k-symbol message m and encodes it with a q-ary LDPC code, resulting in a length-n codeword $C_1 \dots C_n$ over $\mathbb{F}_q$ for a prime q. For a $k \times n$ systematic generator matrix G corresponding to a sparse parity-check matrix H, the resulting codeword is $C = m \cdot G$. The generator matrix can be considered public information. At each step t, the shared secret $S_t = (V_t, \sigma_t)$ is derived by hashing the secret key with the h previously generated tokens, so both parties recompute it from the text itself. It consists of a key value V_t uniform on $\mathcal{R} = \{0, \dots, q-1\}$ and a keyed bijection $\sigma_t : \mathcal{S}_t \to \mathcal{R}$, where $\mathcal{S}_t \subset \mathcal{V}$ is the top-q truncated support of the vocabulary $\mathcal{V}$, so $|\mathcal{S}_t| = q$ and each vertex receives exactly one token. We identify codeword symbols, key values, and colored tokens with the vertices of the $(q-1)$-simplex: the codeword symbol $C_t \in \mathbb{F}_q$ is mapped to vertex C_t, the key value V_t to vertex V_t, and the token $x \in \mathcal{S}_t$ to vertex $\sigma_t(x)$. The channel input vertex at time t is then

$$Z_t = (C_t + V_t) \bmod q. \tag{1}$$

This vertex determines the target location on the simplex toward which the embedding step biases the next-token distribution, while the divergence constraint controls how far the resulting watermarked distribution may deviate from the base model distribution.

Message Embedding SimpleMark constructs a watermarked distribution $Q_t^\star$ by solving a constrained minimization problem which admits a closed-form solution.

For each token, let $\sigma_t(x)$ be the vertex assigned to x, where σ_t is the random bijection shared between the watermarker and detector at time t, and let Z_t be the target vertex defined in (1). We define the

embedding cost $d_t(x) = \mathbf{1}[\sigma_t(x) \neq Z_t]$, which is the discrete metric on the vertex set, since all pairs of vertices are equidistant. The watermarked distribution is then chosen as

$$Q_t^\star \in \arg\min_{P \in \Delta(\mathcal{S}_t)} \sum_{x \in \mathcal{S}_t} P(x) d_t(x) \quad \text{s.t.} \quad D(P, Q_t) \leq \varepsilon, \tag{2}$$

where ε is the selected divergence budget and D is a selectable f-divergence, taken here to be either d_{TV} or D_{KL}. Equivalently, the embedding step moves as much probability mass as permitted by the constraint onto the token coloured Z_t, while keeping the watermarked distribution within distance ε of the base model distribution.

Because the cost is an indicator, the objective in (2) depends on P only through the mass it places on the target vertex. Writing $B = \sum_{x:\sigma_t(x)=Z_t} Q_t(x)$ and $s = \sum_{x:\sigma_t(x)=Z_t} P(x)$, the data-processing inequality gives $D(P, Q_t) \geq D(\mathrm{Bern}(s), \mathrm{Bern}(B))$. Equality is attained by preserving the conditional distributions within the target and non-target bins. The optimum of (2) is therefore the two-scalar tilt

$$Q_t^\star(x) = \begin{cases} \alpha_t \, Q_t(x), & \sigma_t(x) = Z_t, \\ \beta_t \, Q_t(x), & \sigma_t(x) \neq Z_t, \end{cases} \qquad \alpha_t = \frac{s^\star}{B}, \quad \beta_t = \frac{1 - s^\star}{1 - B}, \tag{3}$$

where $s^\star = B + \min(\varepsilon, 1 - B)$ when $D = d_{TV}$, and $s^\star$ is the unique root in $[B, 1]$ of $s \log \frac{s}{B} + (1 - s) \log \frac{1-s}{1-B} = \varepsilon$ when $D = D_{KL}$.[1] In certain settings, it is beneficial to limit the tilt towards unwanted tokens by capping α.

Finally, the next token is sampled according to $X_t \sim Q_t^\star(\cdot \mid x_{<t})$. The generated token is appended to the context, and the procedure is repeated for $t = 1, \ldots, n$.

Message Decoding SIMPLEMARK is decoded with a likelihood-based decoder that uses the watermarked channel law.

Given the prompt and received token sequence $x_{1:n}$, the decoder replays the language model to recover the base next-token distributions $Q_t(\cdot) = Q(\cdot \mid x_{<t})$. It also reconstructs the shared key sequence $S_t = (V_t, \sigma_t)$. For each possible code symbol $a \in \{0, \ldots, q-1\}$, the decoder forms the corresponding target vertex $z_t(a) = (a + V_t) \bmod q$. Using this vertex, the decoder evaluates the probability assigned to the observed token x_t by the divergence-constrained embedding distribution:

$$Q_{t,a}^\star(x) = \begin{cases} \dfrac{s_a^\star}{B_a} \, Q_t(x), & \sigma_t(x) = z_t(a), \\ \dfrac{1 - s_a^\star}{1 - B_a} \, Q_t(x), & \sigma_t(x) \neq z_t(a), \end{cases} \qquad B_a = \sum_{x:\sigma_t(x)=z_t(a)} Q_t(x),$$

with $s_a^\star$ obtained from B_a and ε as in the embedding step. This defines the per-symbol log-likelihood $L_t(a) = \log Q_{t,a}^\star(x_t)$.

For a candidate message $\hat{m}$ with codeword $C(m) = (C_1, \ldots, C_n)$, the decoder sums these likelihoods: $\Lambda(\hat{m}) = \sum_{t=1}^n L_t(C_t(\hat{m}))$. The decoded message is the maximum-likelihood codeword, $\hat{M} = \arg\max_{\hat{m}} \Lambda(\hat{m})$. Simply, the decoder selects the message whose codeword is most likely to explain the observed token sequence under the same divergence-constrained channel used for embedding.

In Appendices A and B, we derive more extensive theoretical results about our watermarking scheme, the watermark channel family, encoding and decoding capabilities, embedding capacity, and perplexity. We discuss a few of those results here. We acknowledge the simplicity of our results; deriving rate under a non-memoryless conditional is much more complicated and remains interesting future work. Most of our theoretical results are applicable to all f-divergences compatible with SIMPLEMARK.

3.2 PERFECT DISTORTION-FREENESS

We show that for perfect distortion freeness, that is when the watermarked distribution is equivalent to that of the unwatermarked, the watermark's capacity is zero. This motivates having an ε-budget that is a measurable f-divergence distance from the unwatermarked distribution.

[1] For $\varepsilon \leq \log(1/B)$. For $\varepsilon > \log(1/B)$, the constrained optimum is simply $s^* = 1$.

Proposition 1 (Rate vanishes as $\varepsilon \to 0$)**.** *Let $\mathcal{X}$ and $\mathcal{Z}$ be finite alphabets, let $Q \in \Delta(\mathcal{X})$ be a fixed base distribution, and let P_Z be a fixed distribution on $\mathcal{Z}$. For $\varepsilon \geq 0$, define the feasible set*

$$\mathcal{F}(\varepsilon) := \left\{ \left(P_{X|Z=z}\right)_{z\in\mathcal{Z}} : d_{\mathrm{TV}}(P_{X|Z=z}, Q) \leq \varepsilon \text{ for all } z \in \mathcal{Z} \right\}.$$

Define the constrained capacity $C(\varepsilon) := \sup_{\left(P_{X|Z=z}\right)\in\mathcal{F}(\varepsilon)} I(Z;X)$, where the mutual information is computed under the joint law $P_{Z,X}(z,x) = P_Z(z)P_{X|Z=z}(x)$. Then $\lim_{\varepsilon\to 0} C(\varepsilon) = 0$. In particular, any achievable embedding rate under the uniform per-key TV constraint must approach 0 as $\varepsilon \to 0$.

Proof. We first show $C(0) = 0$. If $\left(P_{X|Z=z}\right)_{z\in\mathcal{Z}} \in \mathcal{F}(0)$, then for every $z \in \mathcal{Z}$, $d_{\mathrm{TV}}(P_{X|Z=z}, Q) = 0$. Since total variation distance vanishes if and only if the two distributions are equal (Keener, 2010), it follows that $P_{X|Z=z} = Q$, for all $z \in \mathcal{Z}$. Hence the conditional law of X does not depend on Z, so X and Z are independent. Therefore $I(Z;X) = 0$. Taking the supremum over all feasible channels shows that $C(0) = 0$.

We now prove that $C(\varepsilon) \to 0$ as $\varepsilon \to 0$. Let $(\varepsilon_n)_{n\geq 1}$ be any sequence with $\varepsilon_n \downarrow 0$. For each n, choose a feasible channel family $\left(P^{(n)}_{X|Z=z}\right)_{z\in\mathcal{Z}} \in \mathcal{F}(\varepsilon_n)$ such that $I_n := I_{P^{(n)}}(Z;X) \geq C(\varepsilon_n) - \frac{1}{n}$. Because $\mathcal{X}$ and $\mathcal{Z}$ are finite, the product of simplices $\prod_{z\in\mathcal{Z}} \Delta(\mathcal{X})$ is compact (Munkres, 1984). Hence, there exists (Rudin, 1964) a subsequence, which we still index by n, such that for every $z \in \mathcal{Z}$, $P^{(n)}_{X|Z=z} \to P^{\star}_{X|Z=z}$ for some limiting channel family $\left(P^{\star}_{X|Z=z}\right)_{z\in\mathcal{Z}}$.

Since each $P^{(n)}$ is feasible, $d_{\mathrm{TV}}\left(P^{(n)}_{X|Z=z}, Q\right) \leq \varepsilon_n$ for all z. Letting $n \to \infty$ gives $d_{\mathrm{TV}}\left(P^{\star}_{X|Z=z}, Q\right) = 0$ for all z, and therefore $P^{\star}_{X|Z=z} = Q$ for all $z \in \mathcal{Z}$. So under the limiting channel, X is independent of Z, and thus $I_{P^\star}(Z;X) = 0$.

Because mutual information is continuous in the joint distribution on finite alphabets (Polyanskiy & Wu, 2025; Cover & Thomas, 2006), we have $I_n = I_{P^{(n)}}(Z;X) \to I_{P^\star}(Z;X) = 0$. But $C(\varepsilon_n) \leq I_n + \frac{1}{n}$, so taking limits yields $\limsup_{n\to\infty} C(\varepsilon_n) \leq 0$. Since mutual information is always nonnegative, $C(\varepsilon_n) \geq 0$, hence $\lim_{n\to\infty} C(\varepsilon_n) = 0$.

Because this holds for every sequence $\varepsilon_n \downarrow 0$, it follows that $\lim_{\varepsilon\to 0} C(\varepsilon) = 0$. This proves the claim. □

Ideally, a watermarked model would sample from the base distribution itself, and each realized key would leave every next-token distribution exactly untouched. Proposition 1 shows this ideal is unachievable in principle: under exact per-key distortion-freeness the token is independent of the target vertex, the channel collapses, and the embedding capacity is zero; that is, no encoder or decoder can recover a single bit. Distortion is therefore not a defect to be engineered away but the price of embedding, and we make that price the design parameter: an explicit ε divergence budget, with Theorem 1 quantifying the capacity it buys.

3.3 Rate Bounds under Uniform Per-Key TV Constraints

Subsequently, we are able to solve for several capacity results, which can be found in Appendix B. We again emphasize that our theoretical results assume a simpler setting, but nonetheless provide meaningful insight to our watermark's capacity. Let ε denote a TV distance budget.

Theorem 1 (Rate under uniform per-key TV constraint (see Theorem 2))**.** *Assume the conditional-memoryless model. Let a sequence of length-n watermarking schemes satisfy the uniform per-key ε-TV constraint, and let the decoder error probability $P_e^{(n)} := \Pr[\widehat{M} \neq M]$ tend to zero as $n \to \infty$. If $R_n := \frac{k}{n}$ denotes the embedded rate in bits per token, then for every n, $R_n \leq \frac{1}{n}\sum_{t=1}^{n} \widetilde{C}_t(\varepsilon) + \frac{1+nR_nP_e^{(n)}}{n}$, where $\widetilde{C}_t(\varepsilon)$ is the universal constrained channel capacity given in the Appendix as Definition 5. Consequently, if $(R_n)_n$ is bounded and $P_e^{(n)} \to 0$, then $\limsup_{n\to\infty} R_n \leq \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \widetilde{C}_t(\varepsilon)$.*

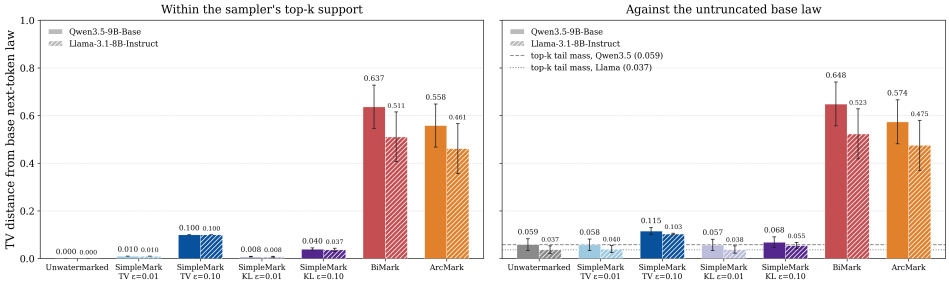

Figure 2: Empirical total variation distance and distinguishing bound from an unwatermarked output and a watermarked output for a baseline, SIMPLEMARK, Bimark, and ARCMARK. The full bars represent Qwen3.5-9B-Base and the dashed bars represent Llama3.1-8B-Instruct.

4 EXPERIMENTAL RESULTS

In this section we demonstrate SIMPLEMARK's efficacy experimentally. We evaluate SIMPLEMARK's reasoning abilities, watermarked token distinguishing advantage, message and bit decoding accuracy, and perplexity. To baseline our performance, we compare our results to ARCMARK (Gilani et al., 2026) and BiMark (Feng et al., 2025). Both watermarks have publicly available implementations, are state of the art for message recovery, and have some definition of distortion freeness the watermark adheres to. We use 300 trials per point unless specified.

4.1 EXPERIMENT PARAMETERS

Setup. We experiment with Qwen3.5-9B-Base (Qwen Team, 2026) and Llama3.1-8B-Instruct (Grattafiori et al., 2024). When applicable, we use the C4-RealNewslike dataset (Raffel et al., 2020) and a temperature of 1.0. Each trial generates 300 tokens and is evaluated at the prefix of $50, 100, \ldots, 300$ tokens. The reasoning experiments use the instruction-tuned Qwen3.5-9B in thinking mode with temperature 0.6. SIMPLEMARK uses a field size of $q = 53$, and subsequently top-53 sampling. The other baselines use top-50 sampling. The ε budget is a per-step TV or KL constraint. We report results for $\varepsilon \in \{0.01, 0.05, 0.1, 0.15\}$ and, where stated, with the per-token likelihood-ratio cap $\alpha_{\max} \in \{500, 1500, 2000\}$. BiMark (Feng et al., 2025) is run with $d = 10$ partition layers, $\tilde{\delta} = 1.0$, context width 2 and top-50 sampling; ArcMark (Gilani et al., 2026) with top-50 sampling, context width 2, $r = 64$ and follow its authors' setup. All schemes see the same prompts. For the key derivation, the encoder and decoder share a fixed integer secret key. At each token position t, the side information is the SHA-256 hash of the secret key concatenated with the ids of the h preceding generated tokens.

4.2 EMPIRICAL MAXIMUM TOKEN DISTINGUISHING ADVANTAGE

We measure the maximum advantage an adversary would have distinguishing between a watermarked and an unwatermarked distribution. As defined in Section 2, total variation distance exactly measures this metric. To empirically estimate TV distance, we prompt the model with 10 prompts and record the model's unwatermarked and watermarked output 10,000 times per prompt, collect the next token output and estimate a distribution. We then calculate the TV distance between the estimated distributions. The results are shown in Figure 2. The outputs have slight noise, as even an unwatermarked output picks up a slight TV distance when estimated based on token outputs.

Interpreting these results, the lower the TV distance, the lower the chance the best possible distinguisher is able to tell whether an output was watermarked or unwatermarked. From our estimated distributions based on the model's outputs, before accounting for noise, for SIMPLEMARK the distinguisher only gains an advantage of at most ε, adhering to the budget, while for ARCMARK, the distinguisher gains a 46–57% advantage and a 51–65% advantage for Bimark. The untruncated law incurs additional TV distance from removing the tail mass outside the top-k. We can take away that in an adversarial setting, such as those used to analyze cryptographic protocols, other current

Scheme (trials)	GSM8K (300)	GPQA-D (198)	LCB (100)	AIME 2024 (30)
3-bit payload				
SimpleMark TV (ε=0.05, $\alpha_{\max}$=1500)	278	112	33	8
SimpleMark KL (ε=0.05, $\alpha_{\max}$=1500)	274	108	40	9
BiMark	270	107	33	11
ArcMark	272	111	35	10
Unwatermarked	272	110	36	8
8-bit payload				
SimpleMark TV (ε=0.05, $\alpha_{\max}$=1500)	275	103	38	9
SimpleMark KL (ε=0.05, $\alpha_{\max}$=1500)	273	102	41	8
BiMark	279	94	33	8
ArcMark	274	109	40	7
Unwatermarked	272	110	36	8

Table 2: Reasoning performance on tasks for Qwen3.5-9B (thinking mode, temperature 0.6) on GSM8K (300 problems), GPQA-Diamond (198 problems), LiveCodeBench (100 problems, graded on public test cases only, so counts are only comparable across schemes) and AIME 2024 (30 problems). We use identical tasks and datasets for every benchmark. We evaluate SimpleMark at two operating points (q=53, TV or KL $\varepsilon = 0.05$, cap $\alpha_{\max}$=1500), BiMark, ArcMark, and no watermark.

watermarking distortion freeness metrics may not be feasible for watermarking schemes to remain secure.

4.3 SIMPLEMARK REASONING PERFORMANCE

We evaluate SIMPLEMARK under several reasoning tasks, GSM8K (Cobbe et al., 2021), GPQA-Diamond (Rein et al., 2023), LiveCodeBench (Jain et al., 2025), and AIME 2024 (Zhang & Math-AI, 2024). We use 300, 198, 100, and 30 problems respectively, with a generation budget of 4,096 tokens for GSM8K, 8,192 for GPQA-Diamond and LiveCodeBench and 16,384 for AIME, and every scheme is run on the identical set of problems with one sample per problem. SIMPLEMARK embeds a 3- or 8-bit message at ε=0.05 (TV) and ε=0.05 (KL), both with $\alpha_{\max}$=1500. On these reasoning traces the cap binds on roughly 91% of tokens.

Table 2 shows that by embedding SIMPLEMARK, the language model's reasoning performance does not degrade. By capping α, we are more likely to avoid sampling undesired tokens that would degrade a model's performance. In Figure 4 we show that we can reliably recover the embedded watermark at a competitive rate under the chosen parameters. The results for Llama3.1-8B-Instruct can be found in Table 3.

4.4 PERPLEXITY

Perplexity is often used as a text quality metric for a watermark. In simple terms, it is how surprised a model is by its output through an exponentiation of the distribution's joint entropy. Perplexity can also be expressed as an exponentiation of the watermarked generating distribution's entropy, plus the KL divergence between the unwatermarked and watermarked distributions. As expected Figure 3 validates that SIMPLEMARK achieves the perplexity closest to that of the unwatermarked distribution under low KL ε-budgets.

4.5 WATERMARK EXTRACTION ACCURACY

We measure message accuracy when recovering our watermark. Bit errors when recovering watermarks can cause the messages to be completely different, so we focus on message accuracy. For medium ε-budgets of $\{0.05, 0.1, 0.15\}$ we generate 300 tokens and use a prefix in increments of 50

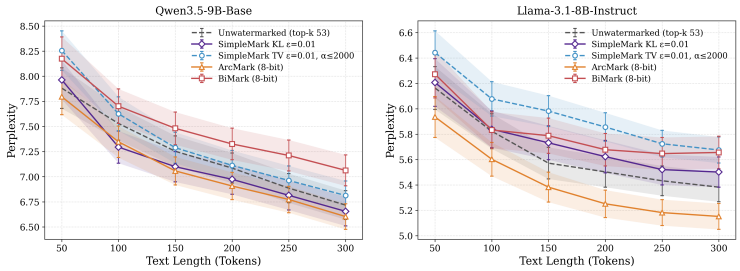

Figure 3: Perplexity results for token outputs ranging from 50-300 tokens for SIMPLEMARK, ARCMARK, and Bimark. A low KL ε-budget SIMPLEMARK is consistently closest to that of the unwatermarked output.

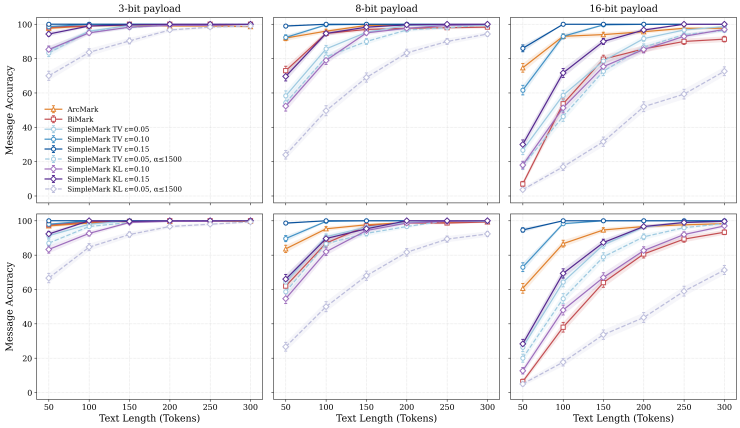

Figure 4: Message recovery for a variable length prefix of a 300 token sequence. For $\varepsilon \in \{0.05, 0.1, 0.15\}$. The SIMPLEMARK decoding capabilities outperform the current SOTA at $\varepsilon = 0.15$, and remain competitive with lower, more practical ε budgets.

tokens up to 300 tokens and attempt to extract the watermark. For lower ε-budgets of $\{0.01, 0.05\}$, we sample from 500–1000 tokens. From Figure 4, at higher ε-budgets SIMPLEMARK's decoding capabilities are competitive with the current state of the art watermarks, which are designed to optimize decoding capabilities and thus have high distortion in token distributions when compared to the unwatermarked law.

Overall, our experiments validate that SIMPLEMARK achieves minimal distortion from the unwatermarked distribution as seen through the distinguishing advantage and perplexity experiments, while retaining a model's reasoning capabilities and competitive extraction accuracy.

5 CONCLUSION

In this work, we studied multi-bit text watermarking as a *constrained communication problem*, highlighting a fundamental tradeoff between payload, reliability, and statistical detectability. Unlike average-key distortion guarantees, our per-key f-divergence formulation directly controls how much each realized watermark distribution can deviate from the base model, in the worst case. We show that exact per-key distortion-freeness forces zero embedding capacity, while nonzero divergence budgets lead to a tunable rate–detectability tradeoff governed by the induced watermark channel. Motivated by this framework, we introduced SIMPLEMARK, which combines error-correcting codes with a simple divergence-constrained probability tilt to obtain a closed-form embedding rule. Empirically, SIMPLEMARK substantially reduces distributional deviation from the base model while maintaining competitive message recovery, perplexity, and reasoning performance. Our results provide an information-theoretic foundation for designing multi-bit watermarks with explicit, tunable statistical-security guarantees.

REFERENCES

André Adelsbach, Stefan Katzenbeisser, and Helmut Veith. Watermarking schemes provably secure against copy and ambiguity attacks. In *Proceedings of the 3rd ACM Workshop on Digital Rights Management*, DRM '03, pp. 111–119, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581137869. doi: 10.1145/947380.947395. URL https://doi.org/10.1145/947380.947395.

Christian Cachin. An information-theoretic model for steganography. *Information and Computation*, 192(1):41–56, 2004. ISSN 0890-5401. doi: https://doi.org/10.1016/j.ic.2004.02.003. URL https://www.sciencedirect.com/science/article/pii/S0890540104000409.

Hongyan Chang, Hamed Hassani, and Reza Shokri. Watermark smoothing attacks against language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4915–4941, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.264. URL https://aclanthology.org/2025.findings-emnlp.264/.

Brian Chen and Gregory W Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information theory*, 47(4):1423–1443, 2001.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. *IACR Cryptol. ePrint Arch.*, pp. 763, 2023. URL https://eprint.iacr.org/2023/763.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Aloni Cohen, Alexander Hoover, and Gabe Schoenbach. Watermarking language models for many adaptive users. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 2583–2601. IEEE, 2025.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006. ISBN 0471241954.

Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. *Digital Watermarking and Steganography*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2 edition, 2007. ISBN 9780080555805.

Xuehao Cui, Ruibo Chen, Yihan Wu, and Heng Huang. Mc^{2}mark: Distortion-free multi-bit watermarking for long messages, 2026. URL https://arxiv.org/abs/2602.14030.

Xiaoyan Feng, He Zhang, Yanjun Zhang, Leo Yu Zhang, and Shirui Pan. Bimark: Unbiased multilayer watermarking for large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Zvyb3WAg03.

Atefeh Gilani, Sajani Vithana, Carol Xuan Long, Oliver Kosut, Lalitha Sankar, and Flavio P Calmon. Arcmark: Distortion-free multi-byte llm watermark via optimal transport. *arXiv preprint arXiv:2602.07235*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Sidong Guo, Sajani Vithana, Atefeh Gilani, Lalitha Sankar, Oliver Kosut, and Flavio P. Calmon. Feedback coding enables inference-time covert agentic communication, 2026. URL https://arxiv.org/abs/2609.24994.

A. Gut. *Probability: A Graduate Course*. Springer Texts in Statistics. Springer New York, 2013. ISBN 9781461447078.

Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. Theoretically grounded framework for LLM watermarking: A distribution-adaptive approach. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=CMmKcHFDKL.

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025.

Ya Jiang, Chuxiong Wu, Massieh Kordi Boroujeny, Brian Mark, and Kai Zeng. StealthInk: A multi-bit and stealthy watermark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27685–27709. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/jiang25j.html.

Ya Jiang, Massieh Kordi Boroujeny, Surender Suresh Kumar, and Kai Zeng. Mirrormark: Generalizable mirrored sampling for multi-bit llm watermarking. *arXiv preprint arXiv:2601.22246*, 2026.

Jonathan Katz and Yehuda Lindell. *Introduction to Modern Cryptography, Second Edition*. Chapman & Hall/CRC, 2nd edition, 2014. ISBN 1466570261.

R.W. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics. Springer New York, 2010. ISBN 9780387938394.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pp. 17061–17084. PMLR, 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

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

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi (Alexis) Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, 6(8): 975–987, 2024. doi: 10.1038/s42256-024-00878-8. URL https://doi.org/10.1038/s42256-024-00878-8.

F.J. MacWilliams and N.J.A. Sloane. *The Theory of Error-Correcting Codes*. North-holland Publishing Company, 2nd edition, 1978.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA, 2nd edition, 2005. ISBN 0521835402. URL https://doi.org/10.1017/CBO9780511813603.

Pierre Moulin. Information-hiding games. In *Digital Watermarking*, pp. 1–12, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-36617-1.

Pierre Moulin. The method of types and its application to information hiding. In *2005 13th European Signal Processing Conference*, pp. 1–1, 2005.

Pierre Moulin and Ralf Koetter. Data-hiding codes. *Proceedings of the IEEE*, 93(12):2083–2126, December 2005.

Pierre Moulin and Joseph A. O'Sullivan. Information-theoretic analysis of information hiding. *IEEE Trans. Inf. Theory*, 49(3):563–593, 2003. doi: 10.1109/TIT.2002.808134. URL https://doi.org/10.1109/TIT.2002.808134.

James R. Munkres. *Elements of Algebraic Topology*. Addison Wesley Publishing Company, 1984.

Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025. ISBN 9781108832908.

Wenjie Qu, Wengrui Zheng, Tianyang Tao, Dong Yin, Yanze Jiang, Zhihua Tian, Wei Zou, Jinyuan Jia, and Jiaheng Zhang. Provably robust multi-bit watermarking for {AI-generated} text. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 201–220, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Sheldon M. Ross. *A First Course in Probability*. Fifth edition, 1998.

W. Rudin. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1964. ISBN 9780070542310. URL https://books.google.com/books?id=yifvAAAAMAAJ.

Jiahao Xu, Rui Hu, Olivera Kotevska, and Zikai Zhang. Xmark: Reliable multi-bit watermarking for llm-generated texts. *arXiv preprint arXiv:2604.05242*, 2026.

Xiaojun Xu, Jinghan Jia, Yuanshun Yao, Yang Liu, and Hang Li. Robust multi-bit text watermark with LLM-based paraphrasers. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DVjkling5x.

KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. Advancing beyond identification: Multi-bit watermark for large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (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. 4031–4055, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.224. URL https://aclanthology.org/2024.naacl-long.224/.

Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. In *Forty-first International Conference on Machine Learning*, 2024.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2024, 2024.

Junlin Zhu, Baizhou Huang, and Xiaojun Wan. Quantilemark: A message-symmetric multi-bit watermark for llms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6790–6806, 2026.

Algorithm 1 Per-key divergence-constrained encoder

Require: Message $m \in \mathcal{M}$, blocklength n, LDPC codebook $\{C_m(1:n)\}_{m\in\mathcal{M}}$ over $\mathbb{F}_q$, secret key K, hash window h, divergence $D \in \{d_{\mathrm{TV}}, D_{KL}\}$, budget ε, base laws $(Q_t)_{t=1}^n$

Ensure: Watermarked token sequence $X_{1:n}$

1: Encode m as the codeword $C_m(1), \ldots, C_m(n)$.

2: **for** $t = 1$ to n **do**

3: Derive the side information from the key and the preceding tokens,

$$(V_t, \sigma_t) \leftarrow \mathrm{Hash}(K,\, X_{t-h:t-1}),$$

where $V_t \in \mathbb{F}_q$ and $\sigma_t : \mathcal{S}_t \to \mathbb{F}_q$ is a keyed bijection on the truncated support.

4: Form the target vertex

$$Z_t \leftarrow (C_m(t) + V_t) \bmod q.$$

5: Compute the base mass on the target vertex,

$$B_t \leftarrow \sum_{x\,:\,\sigma_t(x)=Z_t} Q_t(x),$$

and obtain $s_t^\star$ from (B_t, ε, D): $s_t^\star = B_t + \min(\varepsilon, 1-B_t)$ if $D = d_{\mathrm{TV}}$; the unique root in $[B_t, 1]$ of $s\log\frac{s}{B_t} + (1-s)\log\frac{1-s}{1-B_t} = \varepsilon$ if $D = D_{KL}$. If applicable, cap $s_t^* = \min\{s^*, \alpha_{\max}B\}$

6: Form the closed-form tilt

$$Q_t^\star(x) \leftarrow \begin{cases} \dfrac{s_t^\star}{B_t}\, Q_t(x), & \sigma_t(x) = Z_t, \\ \dfrac{1-s_t^\star}{1-B_t}\, Q_t(x), & \sigma_t(x) \neq Z_t. \end{cases}$$

7: Sample

$$X_t \sim Q_t^\star.$$

8: **end for**

9: Output $X_{1:n}$; the side information $(V_t, \sigma_t)_{t=1}^n$ is recomputable from K and the text itself.

AI USE STATEMENT

In this work, we used generative AI tools for generating the code to run the experiments. We also used generative AI to generate Figure 1 that explains our scheme on a high level.

We have reviewed all AI-assisted work. We take full responsibility for the final content of this work.

A ENCODER AND DECODER ALGORITHMS

SIMPLEMARK's encoding and decoding procedures are described in Section 3.1. In this section, we share the algorithmic forms of Algorithm 1 (encoder) and Algorithm 2 (decoder).

Algorithm 1 describes the encoder: the message is LDPC-encoded once, and each generation step derives the key from the preceding tokens. Subsequently the corresponding target vertex and tilt towards that token are calculated. A token is then sampled from the tilted watermarked distribution.

Algorithm 2 gives the decoder: the decoder starts by recovering the base laws and key sequence from the received text. Next it scores each code symbol under the same f-divergence constrained channel used for embedding, and returns the maximum-likelihood codeword.

B THEORETICAL RESULTS ON CAPACITY OF WATERMARKING CHANNEL

Throughout the appendix, we consider TV distance as the f-divergence to solve for all our theoretical results. As previously stated, some of the theoretical results are limited to simpler settings and most results hold for all f-divergences. We leave additional results to future work.

Also, for simplicity, we assume independent side information and a (conditional) memoryless channel.

Algorithm 2 Per-key divergence-constrained MLE Decoder

Require: Observed tokens $x_{1:n}$, secret key K, hash window h, codebook $\{C_m(1:n)\}_{m\in\mathcal{M}}$, divergence D, budget ε, base model $Q(\cdot \mid \cdot)$

Ensure: Message estimate $\widehat{m}$

1: **for** $t = 1$ to n **do**

2: Recover the base law $Q_t(\cdot) = Q(\cdot \mid x_{<t})$, truncated and renormalized as in encoding, and recompute $(V_t, \sigma_t) \leftarrow \mathrm{Hash}(K,\, x_{t-h:t-1})$.

3: For each symbol $a \in \mathbb{F}_q$, form the candidate vertex and its base mass,

$$z_t(a) = (a + v_t) \bmod q, \qquad B_a \leftarrow \sum_{x\,:\,\sigma_t(x)=z_t(a)} Q_t(x),$$

obtain $s_a^\star$ from (B_a, ε, D) as in Algorithm 1, and store the per-symbol log-likelihood

$$L_t(a) \leftarrow \begin{cases} \log\!\big(\frac{s_a^\star}{B_a}\, Q_t(x_t)\big), & \sigma_t(x_t) = z_t(a), \\ \log\!\big(\frac{1-s_a^\star}{1-B_a}\, Q_t(x_t)\big), & \sigma_t(x_t) \neq z_t(a). \end{cases}$$

4: **end for**

5: **for** each candidate message $m \in \mathcal{M}$ **do**

6: Compute the log-likelihood score

$$\Lambda(m) := \sum_{t=1}^{n} L_t\big(C_m(t)\big).$$

7: **end for**

8: Output any maximizer

$$\widehat{m} \in \arg\max_{m\in\mathcal{M}} \Lambda(m).$$

B.1 Token-Level

Notation and standing assumptions. Let $\mathcal{X} = [1:N]$ denote the token vocabulary. At time t, the base language model induces a next-token distribution

$$Q_t \in \Delta(\mathcal{X}),$$

where $\Delta(\mathcal{X})$ is the probability simplex on $\mathcal{X}$.

As in ArcMark and SimpleMark, the message

$$M \in \mathcal{M} := \{1, \ldots, 2^k\}$$

is encoded into a codeword with an error correcting code

$$C_M(1), \ldots, C_M(n) \in \mathbb{F}_q := \{0, 1, \ldots, q-1\},$$

and the encoder and decoder share side information

$$V_t \sim \mathrm{Unif}(\mathbb{F}_q), \qquad t = 1, \ldots, n.$$

The state variable is the target vertex

$$Z_t := (C_M(t) + V_t) \bmod q \;\in\; \mathcal{Z} := \mathbb{F}_q,$$

identified with the vertices of the $(q-1)$-simplex via the keyed bijection σ_t of Section 3.1.

We assume throughout that:

1. The message M is uniform on $\mathcal{M}$.
2. The side information sequence $V_1, \ldots, V_n$ is i.i.d. and independent of M.

3. Given $(M, V_1, \dots, V_n)$, the encoder generates tokens through a conditional-memoryless channel:
$$P_{X_{1:n}|M,V_{1:n}}(x_{1:n} \mid m, v_{1:n}) = \prod_{t=1}^{n} P_{X_t|Z_t=z_t(m,v_t)}(x_t).$$
4. The decoder observes $(X_{1:n}, V_{1:n})$ and outputs $\widehat{M}$.

Definition 1 (Uniform per-key ε-TV distortion)**.** *Fix $\varepsilon \geq 0$. A watermarking scheme is said to be* uniformly per-key ε-TV distortion-free *if for every time t, every message $m \in \mathcal{M}$, and every realized key value $v \in \mathbb{F}_q$,*
$$d_{\mathrm{TV}}\big(P_{X_t|M=m,V_t=v}, Q_t\big) \leq \varepsilon.$$
Equivalently, since Z_t is a deterministic function of (m, v), the scheme is uniformly per-key ε-TV distortion-free iff
$$d_{\mathrm{TV}}\big(P_{X_t|Z_t=z}, Q_t\big) \leq \varepsilon \qquad \forall z \in \mathcal{Z},\ \forall t.$$

Definition 2 (Feasible per-key channel family)**.** *For each time t and distortion budget $\varepsilon \geq 0$, define*
$$\mathcal{F}_t(\varepsilon) := \left\{ \big(P_{X_t|Z_t=z}\big)_{z\in\mathcal{Z}} : d_{\mathrm{TV}}(P_{X_t|Z_t=z}, Q_t) \leq \varepsilon \ \ \forall z \in \mathcal{Z} \right\}.$$

Definition 3 (Per-key constrained embedding channel)**.** *For each time t, choose a channel family*
$$\big(P^{(\varepsilon)}_{X_t|Z_t=z}\big)_{z\in\mathcal{Z}} \in \mathcal{F}_t(\varepsilon).$$
Given message m and side information v_t, form
$$z_t = z_t(m, v_t) = (C_m(t) + v_t) \bmod q,$$
and sample
$$X_t \sim P^{(\varepsilon)}_{X_t|Z_t=z_t}.$$

Proposition 2 (Uniform per-key distortion-freeness)**.** *If the encoder uses channel families*
$$\big(P^{(\varepsilon)}_{X_t|Z_t=z}\big)_{z\in\mathcal{Z}} \in \mathcal{F}_t(\varepsilon) \qquad \textit{for all } t,$$
then the resulting scheme is uniformly per-key ε-TV distortion-free:
$$d_{\mathrm{TV}}\big(P_{X_t|M=m,V_t=v}, Q_t\big) \leq \varepsilon \qquad \forall t,\ \forall m,\ \forall v.$$

Proof. Fix $t \in [1:n]$, $m \in \mathcal{M}$, and $v \in \mathbb{F}_q$. By construction,
$$Z_t = z_t(m, v) = (C_m(t) + v) \bmod q \in \mathcal{Z}.$$
Conditioned on $(M, V_t) = (m, v)$, the encoder samples X_t according to the conditional law attached to this state:
$$P_{X_t|M=m,V_t=v} = P^{(\varepsilon)}_{X_t|Z_t=z_t(m,v)}.$$
Since
$$\big(P^{(\varepsilon)}_{X_t|Z_t=z}\big)_{z\in\mathcal{Z}} \in \mathcal{F}_t(\varepsilon),$$
the definition of $\mathcal{F}_t(\varepsilon)$ yields
$$d_{\mathrm{TV}}\Big(P^{(\varepsilon)}_{X_t|Z_t=z}, Q_t\Big) \leq \varepsilon \qquad \forall z \in \mathcal{Z}.$$
Applying this to $z = z_t(m, v)$ gives
$$d_{\mathrm{TV}}\Big(P_{X_t|M=m,V_t=v}, Q_t\Big) = d_{\mathrm{TV}}\Big(P^{(\varepsilon)}_{X_t|Z_t=z_t(m,v)}, Q_t\Big) \leq \varepsilon.$$
Because t, m, v were arbitrary, the claim follows uniformly over all times, messages, and realized keys. □

Definition 4 (Single-letter per-time constrained capacity)**.** *For each time t, define*
$$C_t(\varepsilon) := \sup_{\left(P_{X_t|Z_t=z}\right)_{z\in\mathcal{Z}} \in \mathcal{F}_t(\varepsilon)} I(Z_t; X_t),$$
where the mutual information is computed under the joint law
$$P_{Z_t,X_t}(z, x) = P_{Z_t}(z) P_{X_t|Z_t=z}(x).$$

Definition 5 (Universal single-letter constrained capacity)**.** *For each time t, define*

$$\widetilde{C}_t(\varepsilon) := \sup_{\substack{(P_{X_t|Z_t=z})_{z\in\mathcal{Z}}\in\mathcal{F}_t(\varepsilon)\\ P_{Z_t}\in\Delta(\mathcal{Z})}} I(Z_t; X_t),$$

where the mutual information is computed under the joint law

$$P_{Z_t,X_t}(z,x) = P_{Z_t}(z)\, P_{X_t|Z_t=z}(x).$$

Theorem 2 (Converse under uniform per-key TV constraint)**.** *Assume the conditional-memoryless model. Let a sequence of length-n watermarking schemes satisfy the uniform per-key ε-TV constraint, and let the decoder error probability*

$$P_e^{(n)} := \Pr[\widehat{M} \neq M]$$

tend to zero as $n \to \infty$. If

$$R_n := \frac{k}{n}$$

denotes the embedded rate in bits per token, then for every n,

$$R_n \leq \frac{1}{n}\sum_{t=1}^{n} \widetilde{C}_t(\varepsilon) + \frac{1 + nR_nP_e^{(n)}}{n}.$$

Consequently, if $(R_n)_n$ is bounded and $P_e^{(n)} \to 0$, then

$$\limsup_{n\to\infty} R_n \leq \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \widetilde{C}_t(\varepsilon).$$

Proof. Since M is uniform on $\mathcal{M}$ with $|\mathcal{M}| = 2^k$,

$$H(M) = k = nR_n.$$

By Fano's inequality (Cover & Thomas, 2006),

$$H(M \mid X_{1:n}, V_{1:n}) \leq h_2(P_e^{(n)}) + P_e^{(n)} \log(|\mathcal{M}| - 1) \leq 1 + nR_nP_e^{(n)}.$$

Therefore,

$$nR_n = H(M) = I(M; X_{1:n}, V_{1:n}) + H(M \mid X_{1:n}, V_{1:n}),$$

so

$$nR_n \leq I(M; X_{1:n}, V_{1:n}) + 1 + nR_nP_e^{(n)}.$$

Because M and $V_{1:n}$ are independent,

$$I(M; X_{1:n}, V_{1:n}) = I(M; X_{1:n} \mid V_{1:n}).$$

Hence

$$nR_n \leq I(M; X_{1:n} \mid V_{1:n}) + 1 + nR_nP_e^{(n)}.$$

Applying the chain rule gives

$$I(M; X_{1:n} \mid V_{1:n}) = \sum_{t=1}^{n} I(M; X_t \mid X_{1:t-1}, V_{1:n}).$$

Fix t. Let

$$U_t := (X_{1:t-1}, V_{1:n}).$$

Since Z_t is a deterministic function of (M, V_t), and by the conditional-memoryless assumption

$$P_{X_t|M,V_{1:n},X_{1:t-1}} = P_{X_t|Z_t},$$

we have the conditional Markov chain

$$M \longrightarrow Z_t \longrightarrow X_t \qquad \text{given } U_t.$$

Therefore, by data processing (Cover & Thomas, 2006),

$$I(M; X_t \mid U_t) \le I(Z_t; X_t \mid U_t).$$

Thus

$$I(M; X_{1:n} \mid V_{1:n}) \le \sum_{t=1}^{n} I(Z_t; X_t \mid U_t).$$

Now expand the conditional mutual information:

$$I(Z_t; X_t \mid U_t) = \sum_{u} P_{U_t}(u)\, I(Z_t; X_t \mid U_t = u).$$

Fix any u with $P_{U_t}(u) > 0$. Conditioned on $U_t = u$, the channel from Z_t to X_t is still the same single-letter channel family

$$(P_{X_t|Z_t=z})_{z\in\mathcal{Z}},$$

while the input law of Z_t becomes the posterior distribution

$$P_{Z_t|U_t=u}.$$

Because the scheme satisfies the uniform per-key ε-TV constraint, this channel family belongs to $\mathcal{F}_t(\varepsilon)$. Hence, by Definition 5,

$$I(Z_t; X_t \mid U_t = u) \le \widetilde{C}_t(\varepsilon).$$

Since this holds for every u,

$$I(Z_t; X_t \mid U_t) \le \widetilde{C}_t(\varepsilon).$$

Therefore,

$$I(M; X_{1:n} \mid V_{1:n}) \le \sum_{t=1}^{n} \widetilde{C}_t(\varepsilon).$$

Substituting back into the Fano bound yields

$$nR_n \le \sum_{t=1}^{n} \widetilde{C}_t(\varepsilon) + 1 + nR_n P_e^{(n)}.$$

Dividing by n gives

$$R_n \le \frac{1}{n}\sum_{t=1}^{n} \widetilde{C}_t(\varepsilon) + \frac{1 + nR_n P_e^{(n)}}{n}.$$

This proves the finite-n bound.

If $(R_n)_n$ is bounded and $P_e^{(n)} \to 0$, then the additive term on the right tends to 0, so

$$\limsup_{n\to\infty} R_n \le \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \widetilde{C}_t(\varepsilon).$$

□

The following achievability result should be interpreted as an achievability theorem for the induced single-letter channel $P_{X|Z}$, assuming that the encoder can realize codewords $z^n(m) \in \mathcal{Z}^n$ with empirical distribution approaching P_Z. A separate codebook realizability argument is needed to specialize this statement to the constrained SIMPLEMARK or ARCMARK state process $Z_t(m, V_t)$.

Theorem 3 (Achievability in the idealized induced DMC model)**.** *Assume now that:*

1. *$Q_t \equiv Q$ does not depend on t;*
2. *the feasible set is the same at every time, i.e.*

$$\mathcal{F}_t(\varepsilon) \equiv \mathcal{F}(\varepsilon);$$

3. *the encoder uses a fixed channel family*

$$\big(P_{X|Z=z}\big)_{z\in\mathcal{Z}} \in \mathcal{F}(\varepsilon)$$

at every time;

4. *the induced state process $Z_1, \dots, Z_n$ is i.i.d. with marginal P_Z.*

Then any rate

$$R < I(Z;X)$$

is achievable with vanishing block error probability. Consequently, any rate

$$R < C(\varepsilon) := \sup_{\big(P_{X|Z=z}\big)_{z\in\mathcal{Z}}\in\mathcal{F}(\varepsilon)} I(Z;X)$$

is achievable.

Proof. Under the stated assumptions, the induced channel from $Z^n = (Z_1, \dots, Z_n)$ to $X^n = (X_1, \dots, X_n)$ is a discrete memoryless channel with transition law

$$P_{X^n|Z^n}(x^n \mid z^n) = \prod_{t=1}^{n} P_{X|Z}(x_t \mid z_t).$$

Fix any channel family in $\mathcal{F}(\varepsilon)$ and let the input distribution of Z be the induced marginal P_Z. Let $R < I(Z;X)$.

Next, construct a random codebook (MacWilliams & Sloane, 1978) by drawing

$$z^n(1), \dots, z^n(2^{nR})$$

independently, each with i.i.d. components distributed according to P_Z. To transmit message m, send the corresponding codeword $z^n(m)$ through the channel, producing x^n according to

$$\prod_{t=1}^{n} P_{X|Z}(x_t \mid z_t(m)).$$

At the decoder, use jointly typical decoding: declare $\widehat{m}$ to be the unique index such that $(z^n(\widehat{m}), x^n)$ is jointly typical.

By the standard random-coding theorem for discrete memoryless channels (MacWilliams & Sloane, 1978), the average probability of error over the random code ensemble tends to zero as $n \to \infty$ whenever

$$R < I(Z;X).$$

Hence there exists at least one deterministic codebook with block error probability tending to zero at every rate below $I(Z;X)$.

Since the chosen channel family belongs to $\mathcal{F}(\varepsilon)$, the resulting scheme satisfies the per-key TV constraint. Finally, because the argument holds for every feasible channel family, taking the supremum over all

$$\big(P_{X|Z=z}\big)_{z\in\mathcal{Z}} \in \mathcal{F}(\varepsilon)$$

shows that every rate

$$R < C(\varepsilon)$$

is achievable. □

Remark 1 (How to instantiate the constrained channel)**.** *A practical instantiation is to solve, at each time t,*

$$\max_{\big(P_{X_t|Z_t=z}\big)_{z\in\mathcal{Z}}} I(Z_t;X_t) \quad \textit{subject to} \quad d_{\mathrm{TV}}(P_{X_t|Z_t=z}, Q_t) \le \varepsilon \ \ \forall z \in \mathcal{Z}.$$

B.2 Examples for Token-Level

Concrete instantiations of Theorem 2. Theorem 2 states that under the uniform per-key ε-TV constraint, any sequence of watermarking schemes with vanishing decoding error must satisfy

$$R_n \le \frac{1}{n}\sum_{t=1}^{n} \widetilde{C}_t(\varepsilon) + \frac{1 + nR_nP_e^{(n)}}{n},$$

where

$$\widetilde{C}_t(\varepsilon) := \sup_{\substack{(P_{X_t|Z_t=z})_{z\in\mathcal{Z}}\in\mathcal{F}_t(\varepsilon)\\ P_{Z_t}\in\Delta(\mathcal{Z})}} I(Z_t;X_t).$$

Thus, to instantiate the theorem concretely, one fixes a base distribution Q_t, chooses state-conditioned channels

$$\big(P_{X_t|Z_t=z}\big)_{z\in\mathcal{Z}}$$

satisfying

$$d_{\mathrm{TV}}(P_{X_t|Z_t=z}, Q_t) \le \varepsilon \qquad \forall z \in \mathcal{Z},$$

and then computes the resulting mutual information $I(Z_t;X_t)$. This gives an explicit feasible single-letter quantity that can be compared against the converse bound in Theorem 2.

Example 1: binary symmetric perturbation. Let

$$\mathcal{X} = \{0,1\}, \qquad Q(0) = Q(1) = \frac{1}{2},$$

and let the state alphabet be

$$\mathcal{Z} = \{0,1\}, \qquad P_Z(0) = P_Z(1) = \frac{1}{2}.$$

Fix $\varepsilon \in [0, \frac{1}{2}]$ and define the conditional laws

$$P_{X|Z=0}(0) = \frac{1}{2} + \varepsilon, \qquad P_{X|Z=0}(1) = \frac{1}{2} - \varepsilon,$$

and

$$P_{X|Z=1}(0) = \frac{1}{2} - \varepsilon, \qquad P_{X|Z=1}(1) = \frac{1}{2} + \varepsilon.$$

We verify feasibility:

$$d_{\mathrm{TV}}(P_{X|Z=0}, Q) = \frac{1}{2}\Big(\big|\tfrac{1}{2} + \varepsilon - \tfrac{1}{2}\big| + \big|\tfrac{1}{2} - \varepsilon - \tfrac{1}{2}\big|\Big) = \frac{1}{2}(\varepsilon + \varepsilon) = \varepsilon,$$

and similarly

$$d_{\mathrm{TV}}(P_{X|Z=1}, Q) = \varepsilon.$$

Hence this channel family is feasible under the uniform per-key ε-TV constraint.

Now compute the induced marginal on X:

$$P_X(0) = \frac{1}{2}\Big(\frac{1}{2} + \varepsilon\Big) + \frac{1}{2}\Big(\frac{1}{2} - \varepsilon\Big) = \frac{1}{2},$$

and similarly

$$P_X(1) = \frac{1}{2}.$$

Thus $X \sim \mathrm{Bernoulli}(\frac{1}{2})$ and

$$H(X) = 1.$$

Moreover,

$$H(X|Z=0) = h_2\Big(\frac{1}{2} + \varepsilon\Big), \qquad H(X|Z=1) = h_2\Big(\frac{1}{2} + \varepsilon\Big),$$

so

$$H(X|Z) = h_2\Big(\frac{1}{2} + \varepsilon\Big).$$

Therefore

$$I(Z;X) = H(X) - H(X|Z) = 1 - h_2\Big(\frac{1}{2} + \varepsilon\Big).$$

Consequently,

$$\widetilde{C}(\varepsilon) \geq 1 - h_2\Big(\frac{1}{2} + \varepsilon\Big).$$

This gives a completely explicit feasible instantiation of the single-letter quantity appearing in Theorem 2.

Numerical specialization of Example 1. If $\varepsilon = 0.1$, then

$$I(Z;X) = 1 - h_2(0.6) \approx 0.0290 \text{ bits/token.}$$

If $\varepsilon = 0.2$, then

$$I(Z;X) = 1 - h_2(0.7) \approx 0.1187 \text{ bits/token.}$$

Thus, even in this simplest binary example, the theorem predicts that as the allowed per-key distortion budget increases, the admissible embedding rate also increases.

Small-ε expansion for Example 1. Using the Taylor expansion of binary entropy around $\frac{1}{2}$,

$$h_2\Big(\frac{1}{2} + \varepsilon\Big) = 1 - \frac{2}{\ln 2}\varepsilon^2 + o(\varepsilon^2),$$

we obtain

$$I(Z;X) = 1 - h_2\Big(\frac{1}{2} + \varepsilon\Big) = \frac{2}{\ln 2}\varepsilon^2 + o(\varepsilon^2).$$

Hence in this binary symmetric perturbation example, the mutual information scales quadratically in ε for small ε.

Example 2: exact distortion-freeness forces zero information. Now set $\varepsilon = 0$. The per-key TV constraint becomes

$$d_{\mathrm{TV}}(P_{X|Z=z}, Q) = 0 \qquad \forall z \in \mathcal{Z}.$$

Since total variation distance vanishes if and only if the distributions are identical, we obtain

$$P_{X|Z=z} = Q \qquad \forall z \in \mathcal{Z}.$$

Therefore the conditional law of X does not depend on Z, so X and Z are independent:

$$P_{X,Z}(x,z) = P_Z(z)Q(x) = P_Z(z)P_X(x).$$

Hence

$$I(Z;X) = 0.$$

Thus

$$\widetilde{C}_t(0) = 0 \qquad \forall t,$$

and Theorem 2 implies that any asymptotically reliable watermarking scheme under exact per-key distortion-freeness must satisfy

$$\limsup_{n\to\infty} R_n \leq 0.$$

This concretely illustrates the paper's statement that exact per-key distortion-freeness forces zero embedding capacity.

Example 3: ternary perturbation. Let

$$\mathcal{X} = \{1,2,3\}, \qquad Q(1) = Q(2) = Q(3) = \frac{1}{3},$$

and again let

$$\mathcal{Z} = \{0,1\}, \qquad P_Z(0) = P_Z(1) = \frac{1}{2}.$$

Fix $\varepsilon \in [0, \frac{1}{3}]$ and define

$$P_{X|Z=0} = \left(\frac{1}{3} + \varepsilon, \frac{1}{3}, \frac{1}{3} - \varepsilon\right),$$

$$P_{X|Z=1} = \left(\frac{1}{3} - \varepsilon, \frac{1}{3}, \frac{1}{3} + \varepsilon\right).$$

We again verify the TV constraint:

$$d_{\mathrm{TV}}(P_{X|Z=0}, Q) = \frac{1}{2}\left(\left|\frac{1}{3} + \varepsilon - \frac{1}{3}\right| + \left|\frac{1}{3} - \frac{1}{3}\right| + \left|\frac{1}{3} - \varepsilon - \frac{1}{3}\right|\right) = \frac{1}{2}(\varepsilon + 0 + \varepsilon) = \varepsilon,$$

and similarly

$$d_{\mathrm{TV}}(P_{X|Z=1}, Q) = \varepsilon.$$

Thus the channel family is feasible.

Now compute the output marginal:

$$P_X = \frac{1}{2}P_{X|Z=0} + \frac{1}{2}P_{X|Z=1} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) = Q.$$

Hence

$$H(X) = \log_2 3.$$

Also,

$$H(X|Z=0) = H(X|Z=1) = H\left(\frac{1}{3} + \varepsilon, \frac{1}{3}, \frac{1}{3} - \varepsilon\right),$$

so

$$H(X|Z) = H\left(\frac{1}{3} + \varepsilon, \frac{1}{3}, \frac{1}{3} - \varepsilon\right).$$

Therefore

$$I(Z;X) = \log_2 3 - H\left(\frac{1}{3} + \varepsilon, \frac{1}{3}, \frac{1}{3} - \varepsilon\right).$$

This yields another explicit feasible instantiation of the constrained mutual information appearing in Theorem 2.

Example 4: direct SIMPLEMARK instantiation at one time step. Fix a time t, and suppose the base next-token distribution is

$$Q_t = (0.5, 0.3, 0.2)$$

over a three-token vocabulary $\mathcal{X} = \{x_1, x_2, x_3\}$. Suppose also that the relevant state alphabet at time t is binary:

$$\mathcal{Z}_t = \{z_0, z_1\}, \qquad P_{Z_t}(z_0) = P_{Z_t}(z_1) = \frac{1}{2}.$$

For some

$$\delta \in [0, \min\{\varepsilon, 0.2\}],$$

define

$$P_{X_t|Z_t=z_0} = (0.5 + \delta, 0.3, 0.2 - \delta),$$

and

$$P_{X_t|Z_t=z_1} = (0.5 - \delta, 0.3, 0.2 + \delta).$$

Then

$$d_{\mathrm{TV}}(P_{X_t|Z_t=z_0}, Q_t) = \frac{1}{2}\left(|\delta| + 0 + |-\delta|\right) = \delta \le \varepsilon,$$

and similarly

$$d_{\mathrm{TV}}(P_{X_t|Z_t=z_1}, Q_t) \le \varepsilon.$$

Hence this is a feasible per-key channel family.

The output marginal is

$$P_{X_t} = \frac{1}{2}P_{X_t|Z_t=z_0} + \frac{1}{2}P_{X_t|Z_t=z_1} = (0.5, 0.3, 0.2) = Q_t.$$

Therefore

$$H(X_t) = H(0.5, 0.3, 0.2).$$

However, the two conditional entropies are not generally equal. Indeed,

$$H(X_t|Z_t) = \frac{1}{2}H(0.5+\delta, 0.3, 0.2-\delta) + \frac{1}{2}H(0.5-\delta, 0.3, 0.2+\delta).$$

Consequently, the correct mutual information is

$$\boxed{I(Z_t; X_t) = H(0.5, 0.3, 0.2) - \frac{1}{2}H(0.5+\delta, 0.3, 0.2-\delta) - \frac{1}{2}H(0.5-\delta, 0.3, 0.2+\delta).}$$

Thus, for this concrete choice of feasible perturbation, Theorem 2 yields the explicit converse

$$R_n \le \frac{1}{n}\sum_{t=1}^{n} \widetilde{C}_t(\varepsilon) + \frac{1 + nR_nP_e^{(n)}}{n},$$

with

$$\widetilde{C}_t(\varepsilon) \ge H(0.5, 0.3, 0.2) - \frac{1}{2}H(0.5+\delta, 0.3, 0.2-\delta) - \frac{1}{2}H(0.5-\delta, 0.3, 0.2+\delta).$$

Example 5: using the actual SIMPLEMARK state variable. Recall that the state at time t is

$$Z_t = (C_M(t) + V_t) \bmod q,$$

where $C_M(t) \in \mathbb{F}_q$ is the code symbol at time t and $V_t \sim \mathrm{Unif}(\mathbb{F}_q)$ is the secret key value. To obtain a concrete toy instantiation, take $q = 2$, so $Z_t = (C_M(t) + V_t) \bmod 2$ takes values in $\mathcal{Z} = \{0, 1\}$, the two vertices of the 1-simplex; write $z_0 := 0$ and $z_1 := 1$. If the base distribution is

$$Q_t = (0.5, 0.3, 0.2),$$

we may choose the feasible channel family

$$P_{X_t|Z_t=z_0} = (0.5+\delta, 0.3, 0.2-\delta), \qquad P_{X_t|Z_t=z_1} = (0.5-\delta, 0.3, 0.2+\delta),$$

where $\delta \le \varepsilon$. Then $d_{\mathrm{TV}}(P_{X_t|Z_t=z_i}, Q_t) \le \varepsilon$ for $i \in \{0, 1\}$, so this is a valid per-key instantiation in the SIMPLEMARK notation.

If $P_{Z_t}(z_0) = P_{Z_t}(z_1) = 1/2$, then the output marginal is again

$$P_{X_t} = \frac{1}{2}P_{X_t|Z_t=z_0} + \frac{1}{2}P_{X_t|Z_t=z_1} = Q_t.$$

Therefore

$$H(X_t) = H(0.5, 0.3, 0.2),$$

while

$$H(X_t|Z_t) = \frac{1}{2}H(0.5+\delta, 0.3, 0.2-\delta) + \frac{1}{2}H(0.5-\delta, 0.3, 0.2+\delta).$$

Hence the correct per-time mutual information is

$$\boxed{I(Z_t; X_t) = H(0.5, 0.3, 0.2) - \frac{1}{2}H(0.5+\delta, 0.3, 0.2-\delta) - \frac{1}{2}H(0.5-\delta, 0.3, 0.2+\delta).}$$

Thus Theorem 2 says that any asymptotically reliable scheme built from such per-time perturbations must satisfy

$$\limsup_{n\to\infty} R_n \le \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \widetilde{C}_t(\varepsilon),$$

and the displayed quantity above gives an explicit feasible single-letter mutual information value that lower bounds $\widetilde{C}_t(\varepsilon)$.

Interpretation. These examples show concretely how Theorem 2 should be used in practice. One fixes a base next-token distribution Q_t, constructs feasible state-conditioned perturbations lying inside the total variation ball of radius ε around Q_t, computes the induced single-letter mutual information $I(Z_t; X_t)$, and then interprets Theorem 2 as saying that no asymptotically reliable watermarking scheme can have rate larger than the time-average of the best such single-letter quantities. In particular, as $\varepsilon \to 0$, every feasible state-conditioned law collapses back to Q_t, and the mutual information necessarily goes to zero.

Instantiating R_n and $P_e^{(n)}$ in Theorem 2. Recall that in Theorem 2, if the message set has size

$$|\mathcal{M}| = 2^k,$$

and the watermark spans n time steps, then the embedding rate is

$$R_n = \frac{k}{n}$$

bits per token. Moreover, for a decoder producing an estimate $\widehat{M}$ from the observation $(X_{1:n}, V_{1:n})$, the block error probability is

$$P_e^{(n)} := \Pr[\widehat{M} \neq M].$$

Thus, to instantiate the theorem concretely, one must specify a finite blocklength n, a number of embedded bits k, and a decoding rule.

Example 1: binary toy model with $k = 1$ and $n = 1$. Consider the binary construction

$$\mathcal{X} = \{0, 1\}, \qquad \mathcal{Z} = \{0, 1\}, \qquad P_Z(0) = P_Z(1) = \frac{1}{2},$$

with conditional laws

$$P_{X|Z=0}(0) = \frac{1}{2} + \varepsilon, \qquad P_{X|Z=0}(1) = \frac{1}{2} - \varepsilon,$$

$$P_{X|Z=1}(0) = \frac{1}{2} - \varepsilon, \qquad P_{X|Z=1}(1) = \frac{1}{2} + \varepsilon.$$

Suppose we embed one bit, so that

$$M \in \{0, 1\}, \qquad k = 1,$$

and we use a single time step:

$$n = 1.$$

Then

$$R_1 = \frac{k}{n} = 1$$

bit per token.

Assume that the state is identified with the message, i.e. $Z = M$, and let the decoder be the maximum-likelihood decoder

$$\widehat{M}(x) = \begin{cases} 0, & x = 0, \\ 1, & x = 1. \end{cases}$$

Then the block error probability is

$$P_e^{(1)} = \Pr[\widehat{M} \neq M].$$

Since the decoder makes an error exactly when the output flips relative to the more likely symbol under the true message,

$$P_e^{(1)} = \frac{1}{2} \Pr[\widehat{M} \neq 0 \mid M = 0] + \frac{1}{2} \Pr[\widehat{M} \neq 1 \mid M = 1].$$

Now

$$\Pr[\widehat{M} \neq 0 \mid M = 0] = \Pr[X = 1 \mid Z = 0] = \frac{1}{2} - \varepsilon,$$

and similarly

$$\Pr[\widehat{M} \neq 1 \mid M = 1] = \Pr[X = 0 \mid Z = 1] = \frac{1}{2} - \varepsilon.$$

Therefore

$$P_e^{(1)} = \frac{1}{2} - \varepsilon.$$

So in this one-shot example,

$$R_1 = 1, \qquad P_e^{(1)} = \frac{1}{2} - \varepsilon.$$

This illustrates that at blocklength 1, the decoder error remains large unless ε is quite large.

Example 2: repetition over n time steps with $k = 1$. Now use the same binary channel at every time step, but repeat the same embedded bit across n independent uses. Thus

$$M \in \{0, 1\}, \qquad k = 1, \qquad R_n = \frac{1}{n}.$$

Conditioned on $M = 0$, the outputs $X_1, \ldots, X_n$ are i.i.d. with

$$\Pr[X_t = 0 \mid M = 0] = \frac{1}{2} + \varepsilon, \qquad \Pr[X_t = 1 \mid M = 0] = \frac{1}{2} - \varepsilon.$$

Conditioned on $M = 1$, the outputs are i.i.d. with the roles of 0 and 1 reversed.

Let the decoder use majority vote:

$$\widehat{M} = \begin{cases} 0, & \sum_{t=1}^{n} \mathbf{1}\{X_t = 1\} < \frac{n}{2}, \\ 1, & \sum_{t=1}^{n} \mathbf{1}\{X_t = 1\} > \frac{n}{2}. \end{cases}$$

(Without loss of generality, assume that n is odd.)

Then the block error probability is

$$P_e^{(n)} = \Pr[\widehat{M} \neq M].$$

By symmetry,

$$P_e^{(n)} = \Pr\left[\sum_{t=1}^{n} \mathbf{1}\{X_t = 1\} \geq \frac{n}{2} \,\middle|\, M = 0 \right].$$

Since under $M = 0$,

$$\sum_{t=1}^{n} \mathbf{1}\{X_t = 1\} \sim \mathrm{Bin}\left(n, \frac{1}{2} - \varepsilon\right),$$

we obtain

$$P_e^{(n)} = \sum_{j=\lceil n/2 \rceil}^{n} \binom{n}{j} \left(\frac{1}{2} - \varepsilon\right)^j \left(\frac{1}{2} + \varepsilon\right)^{n-j}.$$

Thus in this repetition construction,

$$R_n = \frac{1}{n},$$

while

$$P_e^{(n)} = \sum_{j=\lceil n/2 \rceil}^{n} \binom{n}{j} \left(\frac{1}{2} - \varepsilon\right)^j \left(\frac{1}{2} + \varepsilon\right)^{n-j}.$$

As $n \to \infty$, this error probability tends to zero for every fixed $\varepsilon > 0$, whereas the rate $R_n = 1/n$ also tends to zero. This is consistent with Theorem 2.

Example 3: four messages over two time steps. Suppose now that we wish to embed

$$k = 2$$

bits using

$$n = 2$$

time steps. Then the message set has size

$$|\mathcal{M}| = 2^2 = 4,$$

for example

$$\mathcal{M} = \{00, 01, 10, 11\}.$$

The embedding rate is therefore

$$R_2 = \frac{k}{n} = \frac{2}{2} = 1$$

bit per token.

To make this concrete, suppose at each time t the encoder uses the binary perturbation channel from Example 1, and the two message bits are sent independently across the two time steps:

$$M = (M_1, M_2), \qquad X_t \sim P_{X|Z=M_t}.$$

Let the decoder decode each coordinate separately via

$$\widehat{M}_t = \begin{cases} 0, & X_t = 0, \\ 1, & X_t = 1. \end{cases}$$

Then the overall decoded message is

$$\widehat{M} = (\widehat{M}_1, \widehat{M}_2).$$

Since an error occurs if at least one of the two coordinates is decoded incorrectly, and each coordinate has error probability

$$\frac{1}{2} - \varepsilon,$$

the block error probability is

$$P_e^{(2)} = 1 - \Pr[\widehat{M}_1 = M_1 \text{ and } \widehat{M}_2 = M_2] = 1 - \left(\frac{1}{2} + \varepsilon\right)^2.$$

Equivalently,

$$P_e^{(2)} = 1 - \left(1 - \left(\frac{1}{2} - \varepsilon\right)\right)^2.$$

So in this example,

$$R_2 = 1, \qquad P_e^{(2)} = 1 - \left(\frac{1}{2} + \varepsilon\right)^2.$$

Example 4: general finite-blocklength instantiation. More generally, if one chooses:

- a blocklength n,
- a message set $\mathcal{M}$ with $|\mathcal{M}| = 2^k$,
- a feasible family of per-time channels
$$\left(P_{X_t|Z_t=z}\right)_{z\in\mathcal{Z}},$$
- and a decoder
$$\widehat{M} = \widehat{M}(X_{1:n}, V_{1:n}),$$

then the two quantities in Theorem 2 are instantiated as

$$R_n = \frac{k}{n}$$

and

$$P_e^{(n)} = \Pr[\widehat{M}(X_{1:n}, V_{1:n}) \neq M].$$

For instance, if $k = 100$ bits are embedded over $n = 1000$ tokens, then

$$R_n = \frac{100}{1000} = 0.1$$

bits per token. If the chosen encoder–decoder pair achieves a block error probability of, say,

$$P_e^{(n)} = 10^{-3},$$

then the finite-n converse in Theorem 2 becomes

$$0.1 \leq \frac{1}{1000} \sum_{t=1}^{1000} \widetilde{C}_t(\varepsilon) + \frac{1 + 1000 \cdot 0.1 \cdot 10^{-3}}{1000}.$$

That is,

$$0.1 \le \frac{1}{1000}\sum_{t=1}^{1000}\widetilde{C}_t(\varepsilon) + 0.0011.$$

Equivalently,

$$\frac{1}{1000}\sum_{t=1}^{1000}\widetilde{C}_t(\varepsilon) \ge 0.0989.$$

So this gives a concrete numerical reading of Theorem 2: to support an embedding rate of 0.1 bits per token at block error probability 10^{-3}, the average constrained single-letter mutual information must be at least approximately 0.0989.

Interpretation. In short, R_n is instantiated simply by choosing how many bits k are embedded across how many time steps n, while $P_e^{(n)}$ is instantiated by fixing a decoder and computing its block error probability under the induced joint law of $(M, X_{1:n}, V_{1:n})$. Theorem 2 then relates these concrete finite-blocklength quantities to the average single-letter constrained capacities $\widetilde{C}_t(\varepsilon)$.

B.3 Sequence-Level

Lemma 1 (Sequence-level TV bound from uniform per-step TV control)**.** *Fix a message $m \in \mathcal{M}$ and a realized side-information sequence $v_{1:n} \in \mathbb{F}_q^n$. Let*

$$P_{X_{1:n}}^{(m,v_{1:n})} := P_{X_{1:n}|M=m,V_{1:n}=v_{1:n}}$$

denote the watermarked sequence law, and let

$$Q_{X_{1:n}}^{(n)} := \prod_{t=1}^{n} Q_t$$

denote the reference product law formed from the base next-token distributions. Assume that, conditioned on $(M, V_{1:n}) = (m, v_{1:n})$, the watermarked sequence law factors as

$$P_{X_{1:n}}^{(m,v_{1:n})}(x_{1:n}) = \prod_{t=1}^{n} P_{X_t|M=m,V_t=v_t}(x_t),$$

and that for every $t \in [1:n]$,

$$d_{\mathrm{TV}}\big(P_{X_t|M=m,V_t=v_t},\, Q_t\big) \le \varepsilon_t.$$

Then

$$d_{\mathrm{TV}}\Big(P_{X_{1:n}}^{(m,v_{1:n})},\, Q_{X_{1:n}}^{(n)}\Big) \le 1 - \prod_{t=1}^{n}(1-\varepsilon_t).$$

In particular, if $\varepsilon_t \le \varepsilon$ for all t, then

$$d_{\mathrm{TV}}\Big(P_{X_{1:n}}^{(m,v_{1:n})},\, Q_{X_{1:n}}^{(n)}\Big) \le 1 - (1-\varepsilon)^n \le n\varepsilon.$$

Proof. For each t, write

$$P_t := P_{X_t|M=m,V_t=v_t}, \qquad Q_t := Q_t.$$

By assumption,

$$P_{X_{1:n}}^{(m,v_{1:n})} = \bigotimes_{t=1}^{n} P_t, \qquad Q_{X_{1:n}}^{(n)} = \bigotimes_{t=1}^{n} Q_t.$$

For each t, by the coupling characterization of total variation distance, there exists a coupling

$$\pi_t \in \Pi(P_t, Q_t)$$

of random variables (Y_t, Z_t) such that

$$\Pr[Y_t \ne Z_t] = d_{\mathrm{TV}}(P_t, Q_t) \le \varepsilon_t.$$

Take these couplings independently across t, producing a joint coupling of

$$Y_{1:n} = (Y_1, \dots, Y_n) \sim \bigotimes_{t=1}^{n} P_t$$

and

$$Z_{1:n} = (Z_1, \dots, Z_n) \sim \bigotimes_{t=1}^{n} Q_t.$$

Under this product coupling,

$$\Pr[Y_{1:n} = Z_{1:n}] = \prod_{t=1}^{n} \Pr[Y_t = Z_t] = \prod_{t=1}^{n} \big(1 - \Pr[Y_t \neq Z_t]\big) \geq \prod_{t=1}^{n} (1 - \varepsilon_t).$$

Therefore,

$$\Pr[Y_{1:n} \neq Z_{1:n}] = 1 - \Pr[Y_{1:n} = Z_{1:n}] \leq 1 - \prod_{t=1}^{n} (1 - \varepsilon_t).$$

Applying again the coupling characterization of total variation distance gives

$$d_{\mathrm{TV}}\left(\bigotimes_{t=1}^{n} P_t,\ \bigotimes_{t=1}^{n} Q_t\right) \leq \Pr[Y_{1:n} \neq Z_{1:n}] \leq 1 - \prod_{t=1}^{n} (1 - \varepsilon_t).$$

This proves the first claim.

For the second claim, if $\varepsilon_t \leq \varepsilon$ for all t, then

$$1 - \prod_{t=1}^{n} (1 - \varepsilon_t) \leq 1 - (1 - \varepsilon)^n.$$

Finally, Bernoulli's inequality implies

$$(1 - \varepsilon)^n \geq 1 - n\varepsilon \qquad \text{for } \varepsilon \in [0, 1],$$

and hence

$$1 - (1 - \varepsilon)^n \leq n\varepsilon.$$

This proves the uniform bound. □

Corollary 1 (Uniform sequence-level TV guarantee)**.** *Assume the hypotheses of Lemma 1, and suppose moreover that the scheme is uniformly per-key ε-TV distortion-free, i.e.*

$$d_{\mathrm{TV}}\big(P_{X_t | M=m, V_t=v}, Q_t\big) \leq \varepsilon \qquad \forall t,\ \forall m,\ \forall v.$$

Then for every message m and every realized side-information sequence $v_{1:n}$,

$$d_{\mathrm{TV}}\left(P_{X_{1:n} | M=m, V_{1:n}=v_{1:n}},\ \prod_{t=1}^{n} Q_t\right) \leq 1 - (1 - \varepsilon)^n \leq n\varepsilon.$$

Proof. Apply Lemma 1 with $\varepsilon_t = \varepsilon$ for all t. □

C Distortion-Freeness: Per-Key vs. On-Average

In ARCMARK (Gilani et al., 2026), they consider distortion-freeness, averaged over the secret key. For many applications, this is not an acceptable level of security since LLM watermarking of images/text is done for a particular secret key.

Proposition 3. *The perfect distortion-free condition averaged over the secret key, $\mathbb{E}_S\big[P_{X|S,M=m}\big] = Q_X \quad \forall m$, does not imply per-key perfect distortion-freeness, namely $P_{X|S=s,M=m} = Q_X \quad \forall s, m$.*

Proof. Consider the binary alphabet $\mathcal{X} = \{0, 1\}$ and let $Q_X(0) = Q_X(1) = \frac{1}{2}$. Let the secret key S be uniform on $\{0, 1\}$: $\Pr[S = 0] = \Pr[S = 1] = \frac{1}{2}$. For every message m, define $P_{X|S=0,M=m}(0) = 1, P_{X|S=0,M=m}(1) = 0$, and $P_{X|S=1,M=m}(0) = 0, P_{X|S=1,M=m}(1) = 1$. Then $\mathbb{E}_S\big[P_{X|S,M=m}(0)\big] = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0 = \frac{1}{2} = Q_X(0)$, and similarly $\mathbb{E}_S\big[P_{X|S,M=m}(1)\big] = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1 = \frac{1}{2} = Q_X(1)$. Hence $\mathbb{E}_S\big[P_{X|S,M=m}\big] = Q_X \quad \forall m$. However, $P_{X|S=0,M=m} = \delta_0 \neq Q_X, P_{X|S=1,M=m} = \delta_1 \neq Q_X$. Thus per-key perfect distortion-freeness fails. □

C.1 Separation in Detection

We show that schemes that provide average-key distortion-free guarantees could be strictly less covert against a detector/adversary than the class of uniformly per-key ε-TV distortion-free schemes. In particular, for every $\varepsilon < 1$, there exist average-key distortion-free schemes that are strictly easier to detect than any scheme obeying the uniform per-key ε-TV constraint.

In the proposition below, the detector explicitly receives the secret key. In a real-world deployment, this is (likely) not the case. But we use the proposition to show the separation in detection schemes that satisfy average-key distortion freeness versus per-key distortion freeness.

Proposition 4 (Average-key distortion-freeness can be highly detectable)**.** *Let $\mathcal{X} = \mathcal{V} = \{1, \dots, N\}$, and let $Q(x) = \frac{1}{N}, x \in \mathcal{X}$. Let the secret key $V \sim \mathrm{Unif}(\mathcal{V})$. For each realized key value $v \in \mathcal{V}$, define the keyed output distribution $P_{X|V=v} = \delta_v$. Then the following holds:*

1. *The scheme is perfectly distortion-free on average over the key: $\mathbb{E}_V[P_{X|V}] = Q$.*
2. *The scheme is far from per-key distortion-free. For every $v \in \mathcal{V}$, $d_{\mathrm{TV}}(P_{X|V=v}, Q) = d_{\mathrm{TV}}(\delta_v, Q) = 1 - \frac{1}{N}$.*
3. *If the detector observes both X and the realized key V, then the null law $H_0 : \quad V \sim \mathrm{Unif}(\mathcal{V}), \quad X \sim Q, \quad X \perp V$ and the watermarked law $H_1 : \quad V \sim \mathrm{Unif}(\mathcal{V}), \quad X = V$ have total variation distance $d_{\mathrm{TV}}(P^{H_1}_{X,V}, P^{H_0}_{X,V}) = 1 - \frac{1}{N}$. In particular, the test $\Phi(x, v) = \mathbf{1}\{x = v\}$ satisfies*

$$\mathbb{P}_{H_1}[\Phi(X, V) = 1] = 1, \quad \mathbb{P}_{H_0}[\Phi(X, V) = 1] = \frac{1}{N},$$

and hence has distinguishing advantage $1 - 1/N$.

Consequently, perfect average-key distortion-freeness does not prevent strong detectability once the realized key is used by the detector. Moreover, for every $\varepsilon < 1$, choosing N large enough so that $1 - \frac{1}{N} > \varepsilon$ gives an average-key perfectly distortion-free scheme whose keyed detection advantage exceeds the per-key TV budget ε.

Proof. First, for each $x \in \mathcal{X}$, $\mathbb{E}_V[P_{X|V}(x)] = \frac{1}{N}\sum_{v=1}^{N} \delta_v(x) = \frac{1}{N} = Q(x)$. Therefore $\mathbb{E}_V[P_{X|V}] = Q$, so the scheme is perfectly distortion-free on average over the key.

Next, fix $v \in \mathcal{V}$. Since $P_{X|V=v} = \delta_v$ and Q is uniform, $d_{\mathrm{TV}}(\delta_v, Q) = \frac{1}{2}\sum_{x=1}^{N} |\delta_v(x) - Q(x)|$. The term $x = v$ contributes $\left|1 - \frac{1}{N}\right| = 1 - \frac{1}{N}$, while the remaining $N - 1$ terms contribute $\sum_{x \neq v} \left|0 - \frac{1}{N}\right| = \frac{N-1}{N} = 1 - \frac{1}{N}$. Thus $d_{\mathrm{TV}}(\delta_v, Q) = \frac{1}{2}\left(1 - \frac{1}{N} + 1 - \frac{1}{N}\right) = 1 - \frac{1}{N}$.

It remains to compute the joint distinguishability: Under H_1,

$$P^{H_1}_{X,V}(x, v) = \begin{cases} \frac{1}{N}, & x = v, \\ 0, & x \neq v. \end{cases}$$

Under H_0, X and V are independent uniforms, so $P^{H_0}_{X,V}(x, v) = \frac{1}{N^2} \quad \forall x, v$. Therefore

$$d_{\mathrm{TV}}(P^{H_1}_{X,V}, P^{H_0}_{X,V}) = \frac{1}{2}\sum_{v=1}^{N}\sum_{x=1}^{N} \left|P^{H_1}_{X,V}(x, v) - P^{H_0}_{X,V}(x, v)\right|.$$

For the N diagonal pairs $x = v$, the absolute difference is $\left|\frac{1}{N} - \frac{1}{N^2}\right| = \frac{N-1}{N^2}$. For the $N^2 - N$ off-diagonal pairs $x \neq v$, the absolute difference is $\left|0 - \frac{1}{N^2}\right| = \frac{1}{N^2}$. Hence $d_{\mathrm{TV}}(P^{H_1}_{X,V}, P^{H_0}_{X,V}) = \frac{1}{2}\left[N \cdot \frac{N-1}{N^2} + (N^2 - N) \cdot \frac{1}{N^2}\right]$. Simplifying,

$$d_{\mathrm{TV}}(P^{H_1}_{X,V}, P^{H_0}_{X,V}) = \frac{1}{2}\left[\frac{N-1}{N} + \frac{N-1}{N}\right] = 1 - \frac{1}{N}.$$

Finally, for the test $\Phi(x, v) = \mathbf{1}\{x = v\}$, we have under H_1 that $X = V$ always, so $\mathbb{P}_{H_1}[\Phi(X, V) = 1] = 1$. Under H_0, X and V are independent uniform random variables on $\{1, \dots, N\}$, so $\mathbb{P}_{H_0}[\Phi(X, V) = 1] = \sum_{v=1}^{N} \mathbb{P}[V = v]\mathbb{P}[X = v] = N \cdot \frac{1}{N} \cdot \frac{1}{N} = \frac{1}{N}$. Thus the test has distinguishing advantage $1 - \frac{1}{N}$. This proves the proposition. □

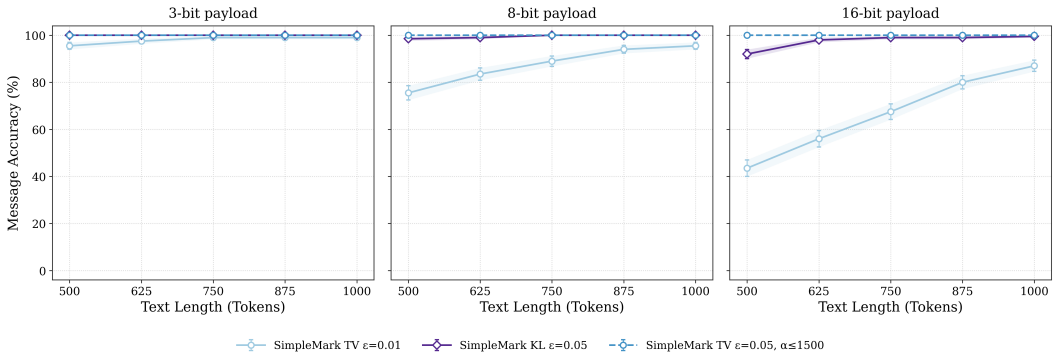

Figure 5: Message recovery for a variable length prefix of a 500-1000 token sequence at various ε budgets.

Scheme	GSM8K (n=300)	GPQA-D (n=198)	LCB (n=300)
3-bit payload			
SimpleMark TV (ε=0.05, $\alpha_{\max}$=1500)	253	58	62
SimpleMark KL (ε=0.05, $\alpha_{\max}$=1500)	249	42	54
BiMark	251	45	58
ArcMark	249	54	65
Unwatermarked	249	56	60
8-bit payload			
SimpleMark TV (ε=0.05, $\alpha_{\max}$=1500)	252	55	60
SimpleMark KL (ε=0.05, $\alpha_{\max}$=1500)	247	61	62
BiMark	248	53	55
ArcMark	259	56	57
Unwatermarked	249	56	60

Table 3: Reasoning performance on several tasks for Llama3.1-8B-Instruct (temperature 0.6) on GSM8K (300 problems×2 independent generations), GPQA-Diamond (198 problems), and LiveCodeBench (300 problems, graded on public test cases only, so counts are only comparable across schemes). We use identical tasks and datasets for every benchmark. We evaluate SimpleMark at two operating points (q=53, TV or KL $\varepsilon = 0.05$, cap $\alpha_{\max}$=1500), BiMark, ArcMark, and no watermark.

D Additional Experiments

We provide additional experimental results. Unless specified in the experiment, the setting is identical to that of Section 4.

D.1 Long-form decoding

We experiment with decoding lower ε-budgets under a 500-1000 token length sequence. SimpleMark's decoder is able to recover the messages quite accurately in this regime. We use the Qwen-3.5-9B-Base model. The results are shown in Figure 5.

D.2 Llama3.1-8B-Instruct Reasoning

We include the Llama3.1-8B-Instruct reasoning results here. All results are consistent with the unwatermarked output in a TV or KL $\varepsilon = 0.05$ decodable regime. AIME was omitted from the table since out of the 2024 and 2025 exams, no model configuration got over 2 problems correct.

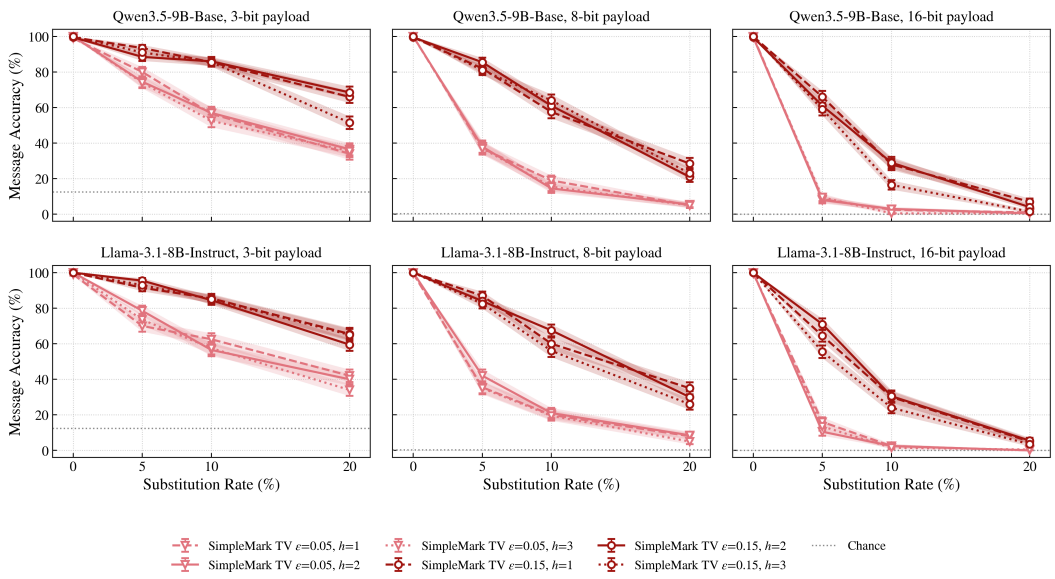

Figure 6: Substitution for a variable percentage of tokens in a 300 token sequence. For $\varepsilon \in \{0.05, 0.15\}$, and context hashing window $h = 2, 3$. SIMPLEMARK is decently robust to substitution edit attacks.

D.3 ROBUSTNESS TO TOKEN SUBSTITUTION

Figure 6 depicts the message decoding accuracy for a 300 token sequence when a $\%$ of the tokens are substituted with a random word in the vocabulary. As expected, the message recovery exponentially degrades as the substitution $\%$ increases.

D.4 ZERO-BIT DETECTION

Figure 7 reports the true positive rate of zero-bit detection as a function of text length at fixed false positive rates of 10^{-2}, 10^{-6}, and 10^{-12}, on Qwen3.5-9B on the top and Llama3.1-8B Instruct on the bottom. Rapid zero-bit detection is not an unambiguous virtue: the keyed detector's power is fueled by the same distributional deviation an adversary can exploit, and the baselines detect quickly precisely because their realized per-token deviations are large (46–65% distinguishing advantage, Figure 2), far in excess of what reliable detection requires. SIMPLEMARK instead caps the keyed per-token distinguishing advantage by ε.

D.5 ABLATION ON DIFFERENT FIELD, TOKEN SAMPLING, AND SIMPLEX SIZE q

Throughout the paper we restricted SIMPLEMARK's simplex, LDPC code field size, and top-k sampling all to 53. In this section we rerun experiments with a smaller parameter of 29. We include message decoding 8, perplexity 9, and GSM8K reasoning 10 baselines with $\alpha_{\max} \in \{200, 500, 1000\}$. All experiments are exactly the same, besides the simplex, LDPC code, and top-k sampling size. We use Qwen-3.5-9B-Base for the reasoning and embed a 16-bit payload.

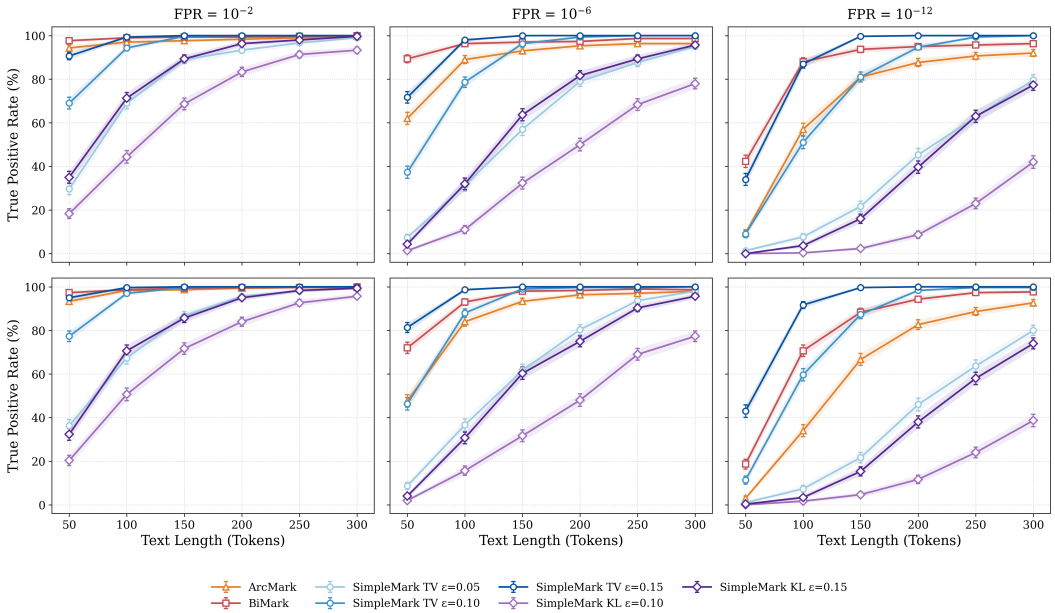

Figure 7: Zero-bit detection for SIMPLEMARK, ARCMARK, and Bimark at different FPR. At higher distortion budgets we see overall improvement on the baselines for detection rate. For lower distortion budgets we stay closer to the unwatermarked distribution, so naturally we see lower detection rates.

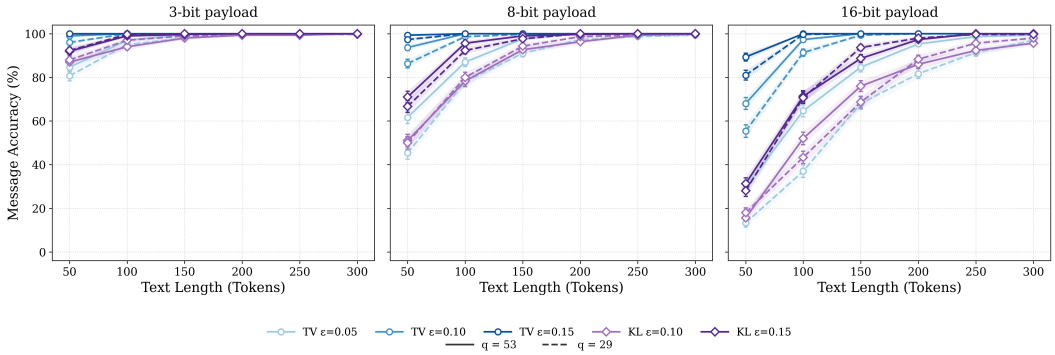

Figure 8: Message recovery for a variable length prefix of a 300 token sequence for $q = \{29, 53\}$ at various ε budgets.

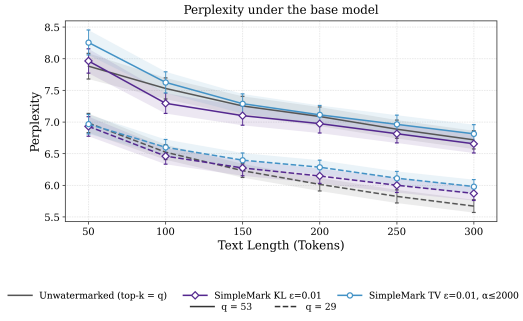

Figure 9: Perplexity with different parameters for $q = \{29, 53\}$. Both remain close to the unwatermarked baselines.

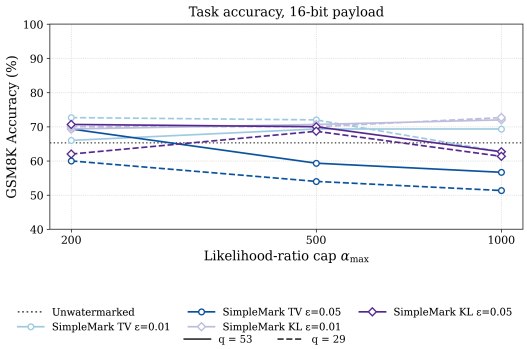

Figure 10: Reasoning for SIMPLEMARK under the different parameter sizes and $\alpha_{\max}$.