# A UNIFIED FRAMEWORK FOR TOKEN-LEVEL WATERMARKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

LLM watermarks allow tracing AI-generated texts by inserting a detectable signal into their generated content. Recent works have proposed a wide range of token-level watermarking algorithms, each with distinct designs. Crucially, there is currently no general and principled formulation for token-level LLM watermarking that could enable deeper understanding of current schemes and ease the design of new variants. In this work, we propose the first unified framework for token-level watermarks based on a principled constrained optimization problem. Our formulation unifies most existing and popular token-level watermarking methods, and explicitly reveals the constraints that each method optimizes. In the process, it highlights an underexplored quality-diversity-power trade-off. Our experimental evaluation validates our framework: watermarking schemes derived from a given constraint consistently maximize detection power with respect to that constraint. At the same time, the framework provides a principled approach for designing novel watermarking schemes tailored to specific requirements, for instance minimizing perplexity of the watermarked text.

## 1 INTRODUCTION

Large Language Model (LLM) token-level watermarking algorithms modify the sampling procedure of the underlying LLM to introduce a detectable signal in the generated text (Kirchenbauer et al., 2023). Importantly, prior works have proposed a range of seemingly unrelated approaches (Kirchenbauer et al., 2023; Dathathri et al., 2024; Aaronson, 2023; Kuditipudi et al., 2024; Christ et al., 2024) to modify the sampling procedure. Usually, given a specific motivation (e.g., distortion-freeness or power), these works first propose a new watermark design, prove that it satisfies some (theoretical) properties, and then experimentally validate it. As a result, it is difficult to compare existing schemes at a fundamental level, relate them to one another, or to design new methods in a principled manner.

**This work:** We address this issue by introducing a unified framework for token-level LLM watermarking. Specifically, we show that constructing a watermarking algorithm can be phrased as a constrained optimization problem: the goal is to maximize the detectability of the inserted watermark signal, while limiting its impact on the quality of the generated text. By formalizing and analyzing this optimization problem, we find that a number of popular schemes (e.g., Red-Green watermarks, AAR/KTH, and SynthID) are the solutions to our optimization problem under specific constraints.

The explicit nature of the constraints in our framework allows us to better understand and relate existing schemes, and provides a principled and practical way to design new and more powerful ones tailored to specific constraints such as quality preservation. Lastly, our framework allows us to theoretically identify and experimentally validate an understudied fundamental trade-off in LLM watermarks: quality-diversity-power, i.e., the trade-off between the quality of the watermarked text, the output diversity under the watermarking scheme, and the overall watermarking signal strength.

**Main contributions** Our main contributions are:

- We formulate token-level LLM watermarking as a general and principled constrained optimization problem that maximizes the detectability under a distortion constraint (Sec. 3).
- We show that our framework captures most token-level watermarking algorithms (Sec. 4.1).
- We instantiate novel schemes that are optimal with respect to a particular objective (Sec. 4.2).



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.

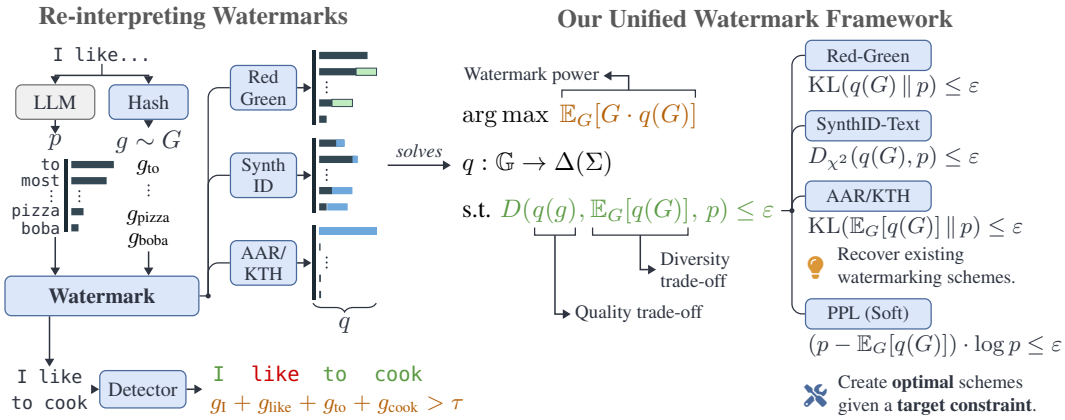



Figure 1: **Overview of Our Framework.** We find that most prior token-level watermarks can be viewed from the following angle (*left*): given a next-token probability distribution $p$ and token scores $g$, pseudorandomly sampled from $G$, they compute a watermarked probability distribution $q(g)$. A text is considered watermarked if the sum of its pseudorandom scores is above a given threshold. Therefore, we frame watermarking as a constrained optimization problem (*middle*): maximizing the expected score while controlling the watermark distortion. In particular, the choice of the constraint balances the watermark quality and diversity. This formulation captures most existing watermarks, and enables the design of novel schemes that are optimal with respect to any given constraint (*right*).

- We empirically show that each scheme (including those from prior works) is optimal with respect to its specific constraint (Sec. 5.1), validate the newly derived schemes across four models and three prompt datasets (Sec. 5.2), evaluate their robustness to text edits (Sec. 5.3), and measure the quality-diversity-power trade-off identified by our framework (Sec. 5.4).

## 2 Background and Related Work

We now introduce the background on token-level LLM watermarks and related work. We defer a taxonomy of watermarks to App. A and the details and pseudocode of each studied scheme to App. D.

**LLM Watermarks** The goal of LLM watermarks is to insert a detectable signal in the text generated by the LLM. At each generation step, watermarks use a hash of the previously generated context to seed a pseudorandom number generator and assign a pseudorandom score to each token in the vocabulary. We call this the *hashing* mechanism. Using these pseudorandom scores, the *sampling mechanism* modifies the original next-token probability distribution provided by the model into a watermarked next-token probability distribution. The next token is then sampled according to the watermarked distribution. Given a text, the *watermark detector* reconstructs the pseudorandom score for each token in the text and computes a p-value. The p-value represents the probability of observing such a result or a more extreme one under the assumption that the sequence was sampled independently of the scores: for human-generated text, the scores are random and the p-value is high, whereas consistently high scores indicate a watermark. Importantly, the detector should be *model-free*, i.e., computing the p-value should not require access to the model and should be computationally efficient. Indeed, practical scenarios assume detection at internet scale, where efficiency is required.

**Scope: Token-Level Watermarks** We focus on *token-level* watermarks, i.e., schemes that, given a fixed hashing mechanism and a fixed score-based detector, embed the watermark solely by modifying the next-token distribution at each step. Such watermarks account for 60% of the 205 papers proposing novel LLM watermarks since 2023 (around 70% when weighted by citations), and include all watermarks currently deployed in industry (Dathathri et al., 2024). Other families, such as rejection-sampling-based (Bahri & Wieting, 2026; Dabiriaghdam & Wang, 2025) and open-weights watermarks (Gu et al., 2024; Block et al., 2025), are outside our scope and are discussed in App. A.

**Watermark Sampling Mechanisms** Prior works have proposed a variety of sampling mechanisms that balance power against impact on quality. Among popular approaches, Kirchenbauer et al. (2023)

Table 1: **Overview of Our Framework.** We summarize which constraints correspond to which watermarks, including prior works and the novel watermarks introduced in Sec. 4. We highlight in light green the novel schemes introduced in this work. We recall that $p$ is the next-token probability distribution of the model, $G$ the random variable representing the pseudorandom token scores, $g$ the sampled token scores (via the hashing mechanism), and $q$ the watermarked probability distribution. For SynthID, we chain $m$ constrained optimization problems corresponding to the $m$ tournament layers and let $q^{(0)} = p$. As we do not extend SynthID, we use its original implementation.

| **Scheme** | Red-Green | AAR/KTH | $\chi^2$ | SynthID | PPL (Hard) | PPL (Soft) |
|---|---|---|---|---|---|---|
| **Constraint** | $\mathrm{KL}(q(g)\|\|p)$ | $\mathrm{KL}(\mathbb{E}_G[q(G)]\|\|p)$ | $\chi^2(q(g), p)$ | $\{\chi^2(q^{(i)}(g), q^{(i-1)}(g))\}_{i=1,\dots,m}$ | $(p - q(g)) \cdot \log p$ | $(p - \mathbb{E}_G[q(G)]) \cdot \log p$ |
| $G$ **Distribution** | Bernoulli | Gumbel | Binomial | Binomial | Binomial | Binomial |
| **Algorithm** | Algorithm 1 | Algorithm 2 | Algorithm 4 | Dathathri et al. (2024) | Algorithm 5 | Algorithm 7 |

biases the logits of the next-token distribution according to binary scores, Dathathri et al. (2024) leverages tournament sampling to favor higher-scoring tokens, Aaronson (2023) deterministically samples the highest-scoring token according to the logarithm of the next-token probability distribution and a Gumbel score, Kuditipudi et al. (2024); Christ et al. (2024) use inverse-transform sampling, and Wu et al. (2024); Chen et al. (2025) reweight the next-token distribution according to the scores' ranking. While these sampling mechanisms are seemingly unrelated, our work shows that most of them can be derived from a unified family of constrained optimization problems (Sec. 4.2).

**Optimal Watermark Design** Several prior works have attempted to design LLM watermarks, or some of their components, from a theoretical perspective. For the hashing mechanism, Hou et al. (2024); Ren et al. (2024); Kodama et al. (2025) propose using a semantics-aware hash to increase the watermark detectability after paraphrasing, and Zhou et al. (2024); Shen et al. (2025) improve the resilience of the hashing mechanism against spoofing attacks. Huang et al. (2023); Li et al. (2025a;b) study the watermark detector to derive the shape of the optimal detector with respect to test power.

More closely related to our work, Wouters (2024); Long et al. (2025) study the sampling mechanism that consists of an exponential tilting of the next-token probability distribution with binary scores (a similar setting to Kirchenbauer et al. (2023)) and find a closed-form formula to compute the watermarked distribution that maximizes power under a log-likelihood constraint. Gloaguen et al. (2026) frame Red-Green watermarks as a constrained optimization problem to derive a watermarking rule for diffusion language models. Tsur et al. (2025) frame watermarking as an optimal transport problem, from which they derive distortion-free schemes that deterministically (according to the pseudorandom scores and $p$) sample each token, recovering the watermarking scheme from Aaronson (2023). Lastly, He et al. (2025) jointly optimize the detector and the sampling mechanism, yielding a powerful yet non-robust watermark. Building on these theoretical insights, they instantiate a distortion-free watermark in which the pseudorandom scores depend on the next-token distribution.

Unlike our framework, existing theoretical frameworks focus solely on finding optimal components for specific LLM watermarks and cannot capture the wide range of existing token-level watermarks.

## 3 Our Framework

In this section, we describe our unified framework for LLM watermarking, as illustrated in Figure 1.

**Setting** We recall the main components of watermarks (Figure 1, left): the hashing mechanism to compute pseudorandom scores, the sampling mechanism that maps the next-token distribution to a watermarked distribution, and the detector that, given a text, assesses whether it is watermarked. Our goal is to design a principled sampling mechanism, given a fixed hashing mechanism and a fixed detector. Let $\Sigma$ be the finite vocabulary, $p \in \Delta(\Sigma)$ be the next-token probability distribution, $G$ the random variable representing the pseudorandom scores with $\mathbb{G} \subset \mathbb{R}^{|\Sigma|}$ as support, and $g \sim G$ a sample from $G$ computed from the context via the hashing mechanism. At every generation step, we want to find a watermarked probability distribution $q(g)$ such that later, without any knowledge of $p$, we can reliably distinguish whether a given sequence of tokens was generated from $p$ or from $q(g)$.

**Optimization Objective** A priori, finding such a distribution $q(g)$ is a complex task. If the detector could depend on $p$ and $q(g)$, the Neyman–Pearson lemma would provide a formula for the best possible detector. Hence, finding $q(g)$ would reduce to maximizing the power of this optimal detector. However, in practice, we want model-free detection (Sec. 2), i.e., the detector should depend on neither $p$ nor $q(g)$. From He et al. (2025), we know that an optimal model-free detector is intractable. In the literature, popular schemes often use the sum of individual token scores as their detection statistic. This suggests that, at each generation step, to maximize the power of the detector one should sample the next token with the highest possible score, i.e., solve $q^*(g) = \arg\max_{q \in \Delta(\Sigma)} \mathbb{E}_{\Omega \sim q}[g_\Omega] = \arg\max_{q \in \Delta(\Sigma)} g \cdot q$. Solving this for all $g \in \mathbb{G}$ amounts to

$$q^* = \arg\max_{q:\mathbb{G}\to\Delta(\Sigma)} \mathbb{E}_G[G \cdot q(G)]. \tag{1}$$

**Adding Constraints** Yet, purely optimizing Equation (1) gives a trivial solution: for all $g$, $q^*(g)$ corresponds to greedy sampling according to the scores $g$, which ignores and thus heavily distorts $p$, leading to low-quality generations. We therefore introduce a constraint to Equation (1) that controls how distortionary the watermark can be. With this, our watermark framework becomes

$$\begin{aligned} q^* = \arg\max_{q:\mathbb{G}\to\Delta(\Sigma)} \quad & \mathbb{E}_G[G \cdot q(G)] \\ \text{s.t. } \forall g \in \mathbb{G}, \quad & D\left(q(g), \mathbb{E}_G[q(G)], p\right) \leq \varepsilon. \end{aligned} \tag{2}$$

The constraint $D\colon \Delta(\Sigma)^3 \to \mathbb{R}$ captures both distortions given a fixed $g$ through $q(g)$, and distortions in expectation through $\mathbb{E}_G[q(G)]$. For instance, $D$ can be a KL constraint $\mathrm{KL}(q(g)||p)$ or a distortion-free constraint $\mathrm{KL}(\mathbb{E}_G[q(G)]||p)$. As we show in Sec. 4, designing a watermarking sampling mechanism thus amounts to choosing a meaningful metric $D$ and then solving Equation (2).

**Hard and Soft Constraints** We define $D$ such that it depends on both $q(g)$ and $\mathbb{E}_G[q(G)]$. We call a constraint on $q(g)$ a *hard constraint*, as the watermark cannot significantly distort the distribution given a fixed $g$. Likewise, we call a constraint on $\mathbb{E}_G[q(G)]$ a *soft constraint*. A single instantiation of $q(g)$ can heavily distort the distribution, but on average, the constraint holds. We find in Sec. 5.4 that soft-constraint watermarks are more powerful and better preserve quality, but exhibit low diversity. In particular, we show in Theorem F.5 that (optimal) watermarks with only soft constraints are necessarily deterministic (i.e., given a fixed prompt, the watermarked LLM always samples the same text). In contrast, hard constraints tend to better preserve diversity at the cost of quality. Although the separation of hard and soft constraints is not strictly necessary in our framework, we find that all existing schemes from Sec. 4.1 can be expressed using only one of the two. This explicit separation provides a novel and more principled perspective on the watermark quality-diversity-power trade-off.

**Penalized Formulation** In practice, solving the constrained optimization problem from Equation (2) at each generation step can be too costly. Hence, for practical reasons, most schemes from Sec. 4 correspond to the penalized version of the optimization problem,

$$\arg\max_{q:\mathbb{G}\to\Delta(\Sigma)} \mathbb{E}_G\left[G \cdot q(G) - \frac{1}{\delta} D\left(q(G), \mathbb{E}_G[q(G)], p\right)\right]. \tag{3}$$

With the penalized version, assuming the existence of an interior point, the shape of the optimal solution $q^*$ is the same as the constrained one: for all $\varepsilon > 0$, there exists $\delta \in \mathbb{R}_+$ such that Equation (2) and Equation (3) have the same solution. Hence, the penalized version corresponds to parametrizing the scheme with the penalty $\delta$ instead of the constraint level $\varepsilon$. In Sec. 4, we parametrize a scheme by $\varepsilon$ if we solve the constrained formulation in practice, and by $\delta$ if we solve the penalized formulation.

## 4 Applying our Framework

In this section, we show how our unified framework naturally describes both existing and new schemes, as illustrated in Figure 1 (right). In Sec. 4.1, we show how our framework captures and extends popular schemes from the literature, and in Sec. 4.2, how to instantiate novel schemes tailored to particular requirements. We summarize our main results in Table 1 and defer all proofs to App. F.

### 4.1 Capturing Existing Watermarking Sampling Mechanisms

We now show how popular sampling mechanisms (Kirchenbauer et al., 2023; Aaronson, 2023; Kuditipudi et al., 2024; Dathathri et al., 2024) can be derived and extended using specific constraints.

**Red-Green Watermarks** One of the most prominent schemes is the Red-Green watermark (KGW) from Kirchenbauer et al. (2023). At each step $t$, the Red-Green watermark pseudorandomly partitions the vocabulary into a set of $\gamma|\Sigma|$ green tokens and the remaining red tokens. Then, it adds a constant $\delta > 0$ to the logits of green tokens, effectively boosting the probability of green tokens. The detector then thresholds the number of green tokens in a given text to decide whether that text is watermarked.

To capture Red-Green watermarks, we let the entries of $G$ be correlated Bernoulli random variables, ensuring that $G$ has exactly $\gamma|\Sigma|$ ones. Then, we use the constraint $D_{\text{R-G}}(q(g), \mathbb{E}_G[q(G)], p) := \text{KL}(q(g)||p)$ for all $g \in \mathbb{G}$. Given the penalization constant $\delta > 0$, Theorem F.1 shows that the solution to Equation (3) is

$$\forall g \in \mathbb{G},\ q_{\text{R-G}}(g) \propto p \exp(\delta g), \quad \text{i.e.,} \quad \log q_{\text{R-G}}(g) = \log p + \delta g + \text{const.} \tag{4}$$

This exactly matches the Red-Green watermark: if a token $u$ is green ($g_u = 1$), we add $\delta$ to its logit. Hence, we have shown that Red-Green watermarks can be naturally derived from our framework as solutions to Equation (3) using a (hard) KL constraint, and we find experimentally (Sec. 5.1) that these Red-Green watermarks are indeed Pareto-optimal with respect to the detectability-KL trade-off.

Because Red-Green watermarks only use a hard constraint, i.e., the constraint depends only on $q(g)$ and not on the distribution of $G$, Equation (4) is the solution of Equation (3) with the KL constraint for any distribution of $G$. This allows us to define new schemes, which we then evaluate in App. E.2.

**AAR and KTH Watermarks** Both Aaronson (2023) and Kuditipudi et al. (2024) (EXP-edit) use the Gumbel-max trick to build a distortion-free watermark. They assign i.i.d. Gumbel scores to each token and sample the next token as the argmax of the sum of the Gumbel scores and the token log-probabilities. In expectation over the Gumbel scores, such sampling is distributed exactly as $p$.

To capture AAR, we use i.i.d. Gumbel scores $G$ and, for all $g \in \mathbb{G}$, the soft constraint

$$D_{\text{AAR}}(q(g), \mathbb{E}_G[q(G)], p) := \text{KL}(\mathbb{E}_G[q(G)]||p). \tag{5}$$

Given $\varepsilon = 0$ (i.e., enforcing distortion-freeness), Theorem F.3 gives us the following solution:

$$\forall u \in \Sigma,\ q_{\text{AAR}}(g)_u := \mathbb{1}\{u = \arg\max_{v \in \Sigma}(g_v + \log p_v)\}. \tag{6}$$

This is exactly the AAR/KTH watermark, which our framework recovers when enforcing the distortion-free constraint with Gumbel scores $G$. This also shows that, when using Gumbel scores with sum-based detectors, AAR/KTH are the most powerful distortion-free watermarking schemes.

Importantly, with $D_{\text{AAR}}$ and $\varepsilon > 0$, our framework extends AAR/KTH to a (more powerful) distortionary variant. Indeed, we show in Theorem F.3 that, when using $D_{\text{AAR}}$, we have:

$$q_{\text{AAR}}(g)_u := \mathbb{1}\left\{u = \arg\max_{v \in \Sigma}\left(g_v + \frac{1}{1+\delta}\log p_v\right)\right\}, \tag{7}$$

where $u \in \Sigma$ and $\delta \geq 0$. We evaluate it in Sec. 5 and find it to be a powerful distortionary scheme.

Our framework also generalizes the AAR/KTH mechanism to any distribution of $G$. For instance, when using a log-normal distribution for $G$, the resulting scheme corresponds to the HeavyWater scheme (Tsur et al., 2025), originally derived using optimal transport. Indeed, Theorem F.3 shows that the resulting mechanisms are always pseudo-deterministic but require solving $|\Sigma|$ equations that depend on $G$. With $\varepsilon = 0$, this is exactly the optimal transport problem from Tsur et al. (2025). Hence, setting $\varepsilon \geq 0$ also extends the schemes from Tsur et al. (2025) with a distortionary version.

**SynthID** Another prominent watermarking scheme is SynthID (Dathathri et al., 2024), which uses $m$-layer tournament sampling as a watermark sampling mechanism to favor higher-scoring tokens.

To capture the $m = 1$ layer tournament, we use i.i.d. Bernoulli scores $G$ and a $\chi^2$ distance constraint,

$$D_{\chi^2}(q(g), \mathbb{E}_G[q(G)], p) := \sum_{u \in \Sigma} \frac{(q(g)_u - p_u)^2}{p_u}. \tag{8}$$

For $\varepsilon$ small enough, Theorem F.8 shows that the solution is $q_{1\text{ layer}}(g) := p(1 + (g - p \cdot g))$. This corresponds to the vectorized 1-layer tournament sampling formula derived in Dathathri et al. (2024). For the $m$-layer tournament, Theorem F.8 shows that it corresponds to chaining the optimization problem $m$ times, each time using the $\chi^2$ constraint with respect to the previous solution. This formulation offers a novel, intuitive perspective on why increasing the number of tournament layers $m$ improves the watermark strength: each additional tournament layer reapplies the $\chi^2$ watermark.

**$\chi^2$ Watermark** With Equation (8), we can also extend the 1-layer tournament beyond the SynthID formulation. In particular, with arbitrary $\varepsilon \geq 0$ and given any $g \in \mathbb{G}$, we find in Theorem F.6 that the sampling mechanism corresponding to the $\chi^2$ constraint is given by

$$\forall u \in \Sigma, q^*(g)_u = p_u[1 + \delta(g_u + \mu)]_+, \tag{9}$$

where $\mu \in \mathbb{R}$ is chosen to satisfy the simplex constraint, $\delta \geq 0$, and $[\cdot]_+$ is the positive part function. This offers a new approach to distortionary 1-layer tournament sampling: whereas SynthID only increases power in discrete steps (by adding candidates per tournament), tuning $\delta$ in Equation (9) trades off power and distortion continuously. We evaluate this scheme using binomial scores in Sec. 5.

### 4.2 Designing Optimal Watermarking Schemes Given Specific Constraints

Next, we show how to use our framework to design watermarking sampling mechanisms that are optimal with respect to specific constraints. As long as computing the solution to Equation (2) is efficient, one can choose any constraint on the simplex as $D$, e.g., the total variation distance or the Hellinger distance. To highlight the practicality of our framework, we propose two new schemes, whose constraints specifically target the quality of the watermarked text. We find in Sec. 5.2 that these new schemes indeed have a better detectability-quality trade-off than those from prior works.

**Perplexity As a Distance** The per-token perplexity (PPL) is defined as the exponential of the information entropy. It measures how surprising it is that a given token has been sampled. The perplexity of a sequence of tokens is the average of the individual token perplexities. In the broader NLP literature, and in the watermarking field in particular, perplexity is commonly used as a proxy to measure the quality impact of a watermark (Kirchenbauer et al., 2023; Dathathri et al., 2024). Our framework allows us to directly use the perplexity (measured with the original model distribution $p$) as a constraint. Specifically, we introduce a hard- and a soft-constraint variant, for all $g \in \mathbb{G}$,

$$D_{\text{Hard PPL}} := (p - q(g)) \cdot \log p \quad \text{and} \quad D_{\text{Soft PPL}} := (p - \mathbb{E}_G[q(G)]) \cdot \log p. \tag{10}$$

**Hard PPL Constraint** Given $\varepsilon \geq 0$, the hard PPL sampling mechanism solves

$$q^* = \underset{q:\mathbb{G}\to\Delta(\Sigma)}{\arg\max}\, \mathbb{E}_G[G \cdot q(G)] \quad \text{s.t.} \quad \forall g \in \mathbb{G},\ D_{\text{Hard PPL}}\left(q(g), \mathbb{E}_G[q(G)], p\right) \leq \varepsilon. \tag{11}$$

This linear program has no closed-form solution but can be solved efficiently with standard solvers. Given $g \in \mathbb{G}$, the optimal distribution $q^*(g)$ is almost deterministic: it is the barycenter of two Dirac distributions, hence of support size at most 2. This is intuitive, as greedy sampling minimizes the PPL of a token sequence. Concretely, ordering tokens by their probability $p$, there exists a rank $r > 0$ such that deterministically sampling any token below this rank satisfies the constraint: if the maximum of $g$ is reached in this subset, we sample this token; otherwise, we find the best combination of a token below and a token above the threshold. We find in Sec. 5 that this scheme is powerful and diverse.

**Soft PPL Constraint** With the soft PPL constraint, Theorem F.5 shows that, because we are only constraining $\mathbb{E}_G[q(G)]$, $q^*(G)$ is a Dirac distribution. Specifically, Theorem F.9 shows that

$$\exists \lambda \in \mathbb{R}, \forall u \in \Sigma,\ q^*(g)_u = \underset{v\in\Sigma}{\arg\max}(g_v + \lambda(\varepsilon) \log p_v), \tag{12}$$

where $\lambda(\varepsilon)$ is the unique solution to $\mathbb{E}_G[q^*(G) \cdot \log p] = p \cdot \log p - \varepsilon$. When $G$ consists of i.i.d. Gumbel variables and $\varepsilon = 0$, we have $\lambda(0) = 1$, recovering exactly the AAR/KTH mechanism. In general, the feasible set of distortion-free watermarks is contained in that of soft-PPL-constrained watermarks, but with i.i.d. Gumbel scores the solutions coincide. For other distributions, $\lambda(\varepsilon)$ has no closed form and is found by root-finding, making this watermark more expensive. Yet, as shown in Sec. 5, this scheme achieves the best detectability-quality trade-off among the evaluated schemes.

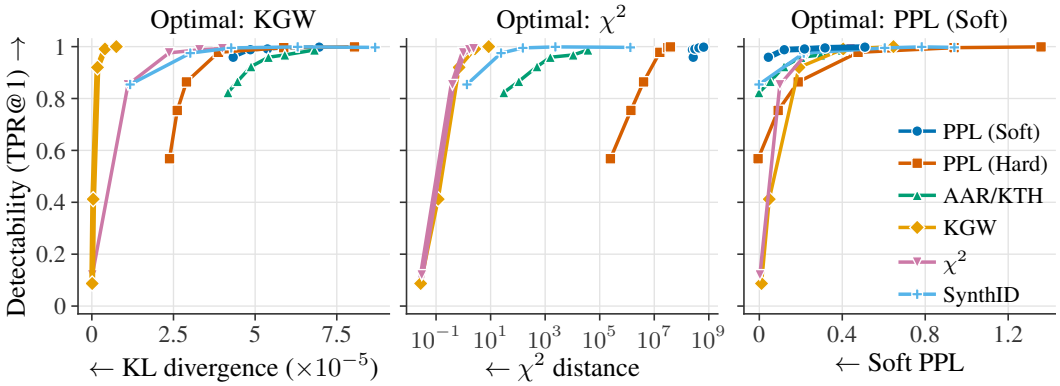



Figure 2: **Comparison of the Detectability–Constraint Trade-off.** We compare the trade-off between watermark detectability (TPR@1) and different constraints (KL divergence (*left*), $\chi^2$ distance (*middle*), and soft PPL (*right*)). We find that, for each constraint, the corresponding scheme (named above each subplot and drawn with a thicker line) has the best detectability-constraint trade-off. Responses are 200-token replies by LLAMA3.1-8B with temperature 0.7 to 1000 ELI5 prompts.

## 5 EVALUATION

We now experimentally validate our framework. In Sec. 5.1, we show that schemes derived from a constraint $D$ are Pareto-optimal with respect to $D$. In Sec. 5.2, we show that both new schemes achieve a state-of-the-art detectability-quality trade-off, and we evaluate their robustness in Sec. 5.3. Lastly, in Sec. 5.4, we measure the quality-diversity-power trade-off from Sec. 3 with a new protocol. We report the full numerical results behind all figures in App. H.

### 5.1 OPTIMALITY WITH RESPECT TO THE CONSTRAINT

**Experimental Setup** Unless stated otherwise, our experiments use LLAMA3.1-8B with temperature 0.7. For each watermark sampling mechanism, we defer the exact hyperparameters (e.g., score distribution, detection algorithm, and hashing mechanism) to App. C. For each watermark and watermark strength, we generate a thousand replies to questions from the ELI5 (Fan et al., 2019) dataset. Each reply is 200 tokens long. For each scheme, we measure the TPR at 1% FPR (we show detectability at other FPRs in App. E.6), and, for each generated token, the constraint (e.g., KL divergence, $\chi^2$ distance) between the original and watermarked next-token probability distributions, which we then average over the sequences. We detail the exact procedure in App. E.1. To estimate $\mathbb{E}_G[q(G)]$ under soft constraints, we use a Monte Carlo approximation based on 128 samples of $G$.

**Results** Figure 2 shows the constraint-detectability trade-off for the KL constraint, the $\chi^2$ constraint, and the soft-PPL constraint. We defer the other constraints from Sec. 5 to App. E.1. Our results show that, as predicted by our framework, Red-Green watermarks are Pareto-optimal for the KL constraint, the $\chi^2$ scheme is optimal for the $\chi^2$ constraint, and the soft-PPL scheme is optimal for the soft-PPL constraint. This result shows that our framework allows the principled construction of powerful schemes with respect to a specific constraint, and correctly captures which specific constraints watermarks from prior work optimize. This experiment directly tests the predictions of our theory on real LLM outputs (App. B). We further compare all schemes under a common quality metric (log PPL) in Sec. 5.2 and under matched score distributions and matched detectors in App. E.2.

### 5.2 DETECTABILITY-QUALITY TRADE-OFF

We next empirically explore the detectability-quality trade-off of all watermarks presented in Sec. 4. Importantly, we show that the extensions of existing schemes (distortionary AAR/KTH and distortionary 1-layer tournament) and the newly proposed PPL-constrained schemes outperform the schemes from prior work. Notably, the soft-PPL scheme achieves a state-of-the-art detectability-quality trade-off, confirming that the soft-PPL constraint is a practical proxy for watermark quality.

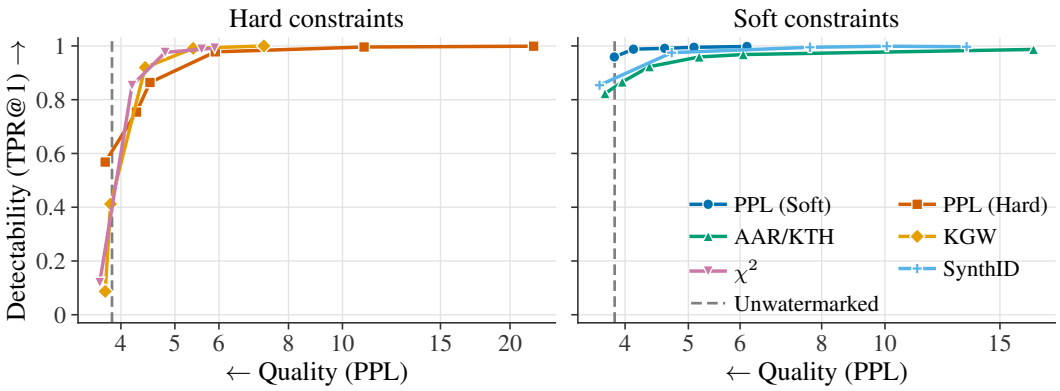



Figure 3: **Comparison of the Detectability–Quality Trade-off.** We compare watermark detectability (TPR@1) and text quality (log PPL) for each constraint and varying $\varepsilon$, with hard (*left*) and soft (*right*) constraints. The dashed line marks the unwatermarked PPL. The setup is the same as in Figure 2.

Table 2: **Robustness Against Sentence-Level Attacks.** We compare the robustness of watermarking schemes against back-translation and paraphrasing attacks using GPT-5.4 MINI. We **bold** the best-performing scheme for each attack. All schemes are evaluated at matched quality (PPL $\approx 3.5$) on replies of 200 to 300 tokens.

| Edit | AAR/KTH | $\chi^2$ | KGW | PPL (Soft) | PPL (Hard) | SynthID |
|---|---|---|---|---|---|---|
| Unmodified | $0.92 \pm 0.02$ | $0.19 \pm 0.02$ | $0.14 \pm 0.02$ | $0.99 \pm 0.01$ | $0.72 \pm 0.03$ | $0.94 \pm 0.02$ |
| Translation | $0.37 \pm 0.03$ | $0.07 \pm 0.02$ | $0.04 \pm 0.01$ | $\mathbf{0.64 \pm 0.03}$ | $0.23 \pm 0.03$ | $0.44 \pm 0.03$ |
| Paraphrasing | $0.05 \pm 0.01$ | $0.03 \pm 0.01$ | $0.01 \pm 0.01$ | $\mathbf{0.13 \pm 0.02}$ | $0.07 \pm 0.01$ | $0.09 \pm 0.02$ |

**Experimental Setup** We use the same experimental setup as in Sec. 5.1. As a proxy for quality, we measure the average text perplexity using QWEN3-30B, and in App. E.4 we use LLM benchmark accuracy. We show examples of watermarked text in App. G, and we repeat this experiment with three other models (MINISTRAL-3-14B, GEMMA4-12B, GPT-OSS-20B) and two further prompt datasets (C4 text completions and low-entropy math prompts from OpenMathInstruct-2) in App. E.3.

**Results** Figure 3 shows the detectability-quality trade-off for all watermarks from Sec. 4, explicitly differentiating between the hard-constraint watermarks (left) and soft-constraint watermarks (right). For hard-constraint watermarks, both the Red-Green watermark and the $\chi^2$ watermark have a similar detectability-quality trade-off, while the hard-PPL watermark exhibits higher detectability in the low-distortion regime. For soft-constraint watermarks, at $\varepsilon = 0$ (i.e., when the log PPL of the watermarked model matches that of the unwatermarked model), the soft-PPL scheme outperforms SynthID, which itself outperforms AAR/KTH. This result is not due to overfitting our quality metric (log PPL): as shown in App. E.4, when using LLM benchmark accuracy as a proxy for quality, the soft-PPL scheme still outperforms AAR/KTH and SynthID. In the distortionary case, where the PPL of the watermarked model exceeds that of the unwatermarked one, the ordering remains similar. These conclusions hold for all additional models and datasets (App. E.3). On low-entropy math prompts, all schemes become less detectable, but the soft-PPL scheme still achieves the best trade-off.

### 5.3 ROBUSTNESS

We evaluate robustness at matched quality (PPL $\approx 3.5$), on replies of 200 to 300 tokens, against back-translation (English $\rightarrow$ French $\rightarrow$ English) and paraphrasing with GPT-5.4 MINI, deferring word-level edits and DIPPER (Krishna et al., 2023) paraphrasing to App. E.5. Table 2 shows that the soft-PPL scheme is the most detectable on unmodified, back-translated, and paraphrased text, and leads under word-level edits and weak DIPPER paraphrasing. However, like all other evaluated watermarks, it loses most of its signal under strong paraphrasing (Zhang et al., 2024). Paraphrasing robustness is usually achieved via the hashing mechanism (e.g., semantic hashing (Hou et al., 2024; Kodama et al., 2025)), which is orthogonal to our focus on sampling.

### 5.4 Impact on Diversity

Although prior work (Gloaguen et al., 2025b) has suggested that watermarks, especially distortion-free schemes, may reduce diversity, this effect has not been rigorously evaluated. We propose a new experimental protocol to quantify it, and find that soft-constrained watermarks substantially reduce diversity, whereas hard-constrained watermarks largely preserve it, as predicted by our framework.

**Measuring Watermark Diversity** As in Sec. 5.2, we use LLAMA3.1-8B with temperature 0.7. As observed in Gloaguen et al. (2025b), watermarks reduce the number of unique outputs given a fixed prompt, so reliably measuring diversity requires sampling the model multiple times with the same prompt. If a watermark reduces diversity, we expect more repetitions across the generated answers. In the limit, AAR/KTH sampling is deterministic given a fixed prompt, so all texts are identical.

We therefore propose the following experimental setup. Given a dataset of prompts $\{x_i\}_i$, for each prompt, we sample the model $N$ times and compute the self-BLEU score (Alihosseini et al., 2019) for that prompt. Self-BLEU scores are between 0 and 1, and a high score means that the generations for a given prompt have similar n-grams. To smooth out the effect of the prompt itself, we then average all self-BLEU scores. In practice, we use 100 prompts from the ELI5 dataset and sample the model $N := 100$ times per prompt, yielding a total of 10,000 samples. Each sample has 200 tokens, and we use trigram self-BLEU with uniform weights.

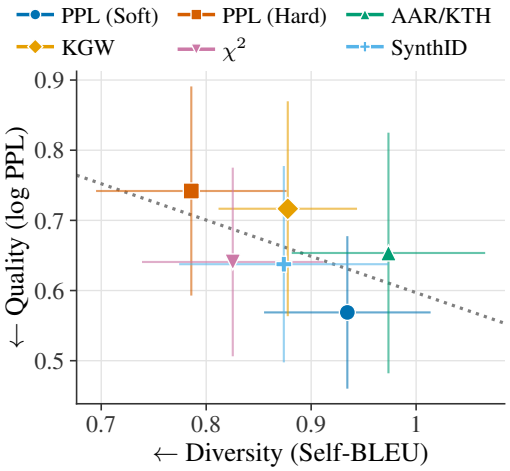



Figure 4: **Diversity–Quality Trade-off at Fixed TPR.** We compare the trade-off between watermark impact on diversity (Self-BLEU) and quality (log PPL) given a fixed TPR@1 of 0.95. A high self-BLEU score corresponds to a low output diversity. Error bars show the standard deviation and the dotted gray line the correlation. We find that, given a fixed detection power, diversity and text quality are negatively correlated.

**Watermark Diversity-Quality Trade-off** Figure 4 shows the impact on diversity and quality of different watermarks given a fixed TPR@1. As predicted by our framework, we find a negative correlation between diversity and quality. In particular, soft-constrained schemes (AAR/KTH and PPL (Soft)) have particularly low diversity (i.e., high self-BLEU scores) but also the highest quality. This confirms the theoretical derivation from Sec. 3: at a fixed power, there is a fundamental trade-off between watermark quality and diversity. More generally, this suggests that to measure the impact of distortion-free watermarks, it is also necessary to evaluate the watermark behavior under repetition of the same prompt (a similar finding to Wu et al. (2026)), rather than assuming that distortion-freeness implies no actual distortion. We discuss the practical implications of this reduced diversity in App. B.

## 6 Conclusion and Limitations

Our work shows that token-level watermarks can be seen as instances of a unified constrained optimization problem that (i) provides a principled way to interpret existing watermarks, (ii) allows us to practically instantiate powerful new schemes tailored to specific watermarking requirements, and (iii) provides a novel perspective on the quality-diversity-power trade-off in LLM watermarking.

**Limitations** Our framework captures token-level watermarks and does not apply to other, less prominent categories (e.g., rejection-sampling-based or open-weights watermarks, App. A). Our optimality results are relative to a fixed hashing mechanism, score distribution, and sum-based detector: we do not jointly optimize the detector, as prior work shows that doing so naively yields non-robust watermarks (He et al., 2025) (App. B). The soft-PPL scheme requires costly per-step root-finding. Like all evaluated watermarks, our schemes are only mildly robust to strong paraphrasing: our framework primary aim is to understand existing schemes, e.g., KGW, AAR, or SynthID.

### LLM USAGE

The paper was written by hand, with AI assistance for polishing the writing and the visuals. For the research, AI assistance was used to implement the experiments and to help with mathematical results.

### ETHICS STATEMENT

This paper presents a unified framework for LLM watermarking, which allows us to better understand existing schemes and design new ones. Better-understood and better-tuned watermarks could support provenance, auditing, and mitigation of large-scale AI-generated text misuse. While LLM watermarks provide statistical guarantees on false positives, they should not be the sole evidence in high-stakes decisions. We also acknowledge potential misuses of watermarks (e.g., tracking user content without their consent), and discuss these implications in App. A.

### REPRODUCIBILITY STATEMENT

All theoretical results are stated with all assumptions and proven in App. F. The pseudocode of every evaluated scheme is in App. D, and all hyperparameters (score distributions, hashing, detection tests, sampling, and quality models) are given in App. C. All datasets and models used are publicly available (App. I), and each additional experiment is detailed in the corresponding section of App. E.

## REFERENCES


Scott Aaronson. Watermarking of large language models. Talk at the Workshop on Large Language Models and Transformers, Simons Institute, UC Berkeley, 2023. URL https://simons.berkeley.edu/talks/scott-aaronson-ut-austin-openai-2023-08-17.

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 90–98, 2019.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Dara Bahri and John Wieting. A watermark for black-box language models. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=6gcHcgGmLo.

Adam Block, Alexander Rakhlin, and Ayush Sekhari. GaussMark: A practical approach for structural watermarking of language models. In *International Conference on Machine Learning*, pp. 4507–4569. PMLR, 2025.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Ruibo Chen, Yihan Wu, Junfeng Guo, and Heng Huang. Improved unbiased watermark for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20587–20601, 2025. doi: 10.18653/v1/2025.acl-long.1005.

Miranda Christ and Sam Gunn. Pseudorandom error-correcting codes. In *Annual International Cryptology Conference (CRYPTO)*, 2024.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 1125–1139. PMLR, 2024.

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

Amirhossein Dabiriaghdam and Lele Wang. Simmark: A robust sentence-level similarity-based watermarking algorithm for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30773–30794, 2025. doi: 10.18653/v1/2025.emnlp-main.1567.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

A. Dvoretzky, A. Wald, and J. Wolfowitz. Elimination of Randomization in Certain Statistical Decision Procedures and Zero-Sum Two-Person Games. *The Annals of Mathematical Statistics*, 22(1):1 – 21, 1951. doi: 10.1214/aoms/1177729689. URL https://doi.org/10.1214/aoms/1177729689.

Jaiden Fairoze, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, and Mingyuan Wang. Publicly-detectable watermarking for language models. *IACR Communications in Cryptology*, 1(4), 2024. doi: 10.62056/ahmpdkp10.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3558–3567, 2019. doi: 10.18653/v1/P19-1346.

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. Three bricks to consolidate watermarks for large language models. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2023. doi: 10.1109/WIFS58808.2023.10374576.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Thibaud Gloaguen, Nikola Jovanović, Robin Staab, and Martin Vechev. Towards watermarking of open-source llms. *arXiv preprint arXiv:2502.10525*, 2025a.

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

Thibaud Gloaguen, Robin Staab, Nikola Jovanović, and Martin Vechev. Watermarking diffusion language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3aBWTYGcaT.

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

Yuxuan Guo, Zhiliang Tian, Yiping Song, Tianlun Liu, Liang Ding, and Dongsheng Li. Context-aware watermark with semantic balanced green-red lists for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 22633–22646. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.1260. URL https://doi.org/10.18653/v1/2024.emnlp-main.1260.

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.

Abe Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. Semstamp: A semantic watermark with paraphrastic robustness for text generation. In *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, 2024.

Baihe Huang, Banghua Zhu, Hanlin Zhu, Jason Lee, Jiantao Jiao, and Michael Jordan. Towards optimal statistical watermarking. In *Socially Responsible Language Modelling Research*, 2023. URL https://openreview.net/forum?id=Fc2FaS9mYJ.

Jiahao Huo, Shuliang Liu, Bin Wang, Junyan Zhang, Yibo Yan, Aiwei Liu, Xuming Hu, and Mingxun Zhou. PMark: Towards robust and distortion-free semantic-level watermarking with channel constraints. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EhDgP69DJG.

Nikola Jovanović, Robin Staab, and Martin Vechev. Watermark stealing in large language models. In *International Conference on Machine Learning*, pp. 22570–22593. PMLR, 2024.

Nikola Jovanović, Ismail Labiad, Tomas Soucek, Martin Vechev, and Pierre Fernandez. Watermarking autoregressive image generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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*, pp. 17061–17084. PMLR, 2023.

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

Shingo Kodama, Haya Diwan, Lucas Rosenblatt, R. Teal Witter, and Niv Cohen. Simkey: A semantically aware key module for watermarking language models, 2025. URL https://arxiv.org/abs/2510.12828.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, 2023.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *TMLR*, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *The Annals of Statistics*, 53(1):322–351, 2025a.

Xingchi Li, Guanxun Li, and Xianyang Zhang. A likelihood based approach for watermark detection. In *The 28th International Conference on Artificial Intelligence and Statistics*, pp. 1675–1683. PMLR, 2025b.

Carol Xuan Long, Dor Tsur, Claudio Mayrink Verdun, Hsiang Hsu, Haim Permuter, and Flavio P Calmon. Optimized couplings for watermarking large language models. In *2025 IEEE International Symposium on Information Theory (ISIT)*, pp. 1–6, 2025. doi: 10.1109/ISIT63088.2025.11195457.

Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. MarkLLM: An open-source toolkit for LLM watermarking. In Delia Irazu Hernandez Farias, Tom Hope, and Manling Li (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 61–71, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-demo.7.

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.

Jie Ren, Han Xu, Yiding Liu, Yingqian Cui, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. A robust semantics-based watermark for large language model against paraphrasing. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 613–625, 2024.

Huanming Shen, Baizhou Huang, and Xiaojun Wan. Enhancing llm watermark resilience against both scrubbing and spoofing attacks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Katherine Thai, Bradley Emi, Elyas Masrour, and Mohit Iyyer. EditLens: Quantifying the extent of AI editing in text. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gOkitaPCfZ.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. In *The Thirteenth International Conference on Learning Representations*, 2025.

Dor Tsur, Carol Xuan Long, Claudio Mayrink Verdun, Sajani Vithana, Hsiang Hsu, Chun-Fu Chen, Haim Permuter, and Flavio P Calmon. Heavywater and simplexwater: Distortion-free llm watermarks for low-entropy distributions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Bram Wouters. Optimizing watermarks for large language models. In *International Conference on Machine Learning*, pp. 53251–53269. PMLR, 2024.

Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. A resilient and accessible distribution-preserving watermark for large language models. In *International Conference on Machine Learning*, pp. 53443–53470. PMLR, 2024.

Yihan Wu, Xuehao Cui, Ruibo Chen, and Heng Huang. Analyzing and evaluating unbiased language model watermark. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6T4LR1oRwA.

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

Junyan Zhang, Shuliang Liu, Aiwei Liu, Yubo Gao, Jungang Li, Xiaojie Gu, and Xuming Hu. Cohemark: A novel sentence-level watermark for enhanced text quality, 2025. URL https://arxiv.org/abs/2504.17309.

Tong Zhou, Xuandong Zhao, Xiaolin Xu, and Shaolei Ren. Bileve: Securing text provenance in large language models against spoofing with bi-level signature. *Advances in Neural Information Processing Systems*, 37:56054–56075, 2024.

# A A TAXONOMY OF LLM WATERMARKS

In this section, we propose an in-depth taxonomy of LLM watermarks and show that token-level watermarking is currently the most prominent, widely used, and practical approach for LLM watermarking. We also discuss the societal implications of watermarking.

**LLM Watermarks** The goal of LLM watermarks is to insert a detectable signal into the text generated by the LLM. Then, given a text, the watermark detector can detect this signal and decide whether this text was generated using the watermarking algorithm. Importantly, unlike generic AI-detection tools (Thai et al., 2026), the watermark detector should provide a theoretical guarantee on its false-positive rate (e.g., by computing statistically sound p-values). At the same time, previous work has identified three main challenges for LLM watermarking: robustness (Kirchenbauer et al., 2024) (i.e., , can the watermark be reliably detected after edits to the generated text), security (Jovanović et al., 2024) (i.e., , can an adversary remove or spoof the watermark signal), and the applicability of watermarks to open-weights models (Gloaguen et al., 2025a) (i.e., , can watermarks be used with open-weights models). These challenges give rise to various categories of watermarking algorithms: token-level watermarks that operate jointly with LLM sampling, rejection-sampling-based watermarking schemes that use rejection sampling to inject the watermark signal, and open-weight watermarks that are embedded directly into the model weights.

**Token-Level Watermarks** Token-level watermarks are the main focus of this work and are described in Sec. 2. Formally, we call a watermark *token-level* if, given a fixed pseudorandom hashing mechanism that assigns a score to each token of the vocabulary and a fixed score-based detector, it embeds the watermark solely by modifying the next-token distribution at each generation step. This is precisely the setting captured by our framework (Sec. 3). Token-level watermarks represent a very general subset of LLM watermarking schemes, and importantly they are the only ones currently deployed in the industry (Dathathri et al., 2024). To quantify their prevalence, we collected all papers mentioning both language models and watermarks, and used GPT-5.4 to identify those proposing a novel watermarking algorithm and to classify each such algorithm as token-level or not. Out of the 205 papers proposing novel LLM watermarks since 2023, 60% propose token-level watermarks, and, weighted by citations, these account for around 70% of all citations. Indeed, token-level watermarks naturally fit LLM sampling, which operates at the token level, making them fast, easy-to-implement, and reliable. At the same time, they are also flexible. Prior works have shown how to use them to improve watermark robustness with semantic-aware watermarking (Guo et al., 2024; Kodama et al., 2025; Ren et al., 2024) and watermark security (Zhou et al., 2024), how to adapt them to diffusion language models (Gloaguen et al., 2026), how to turn them into open-weights watermarks (Gu et al., 2024), and even how to adapt them for image watermarking (Jovanović et al., 2025).

**Rejection-Sampling Watermarks** At a high level, rejection-sampling-based watermarks operate by sampling multiple units of text (usually sentences) in parallel from the LLM, scoring them according to a pseudo-random function, and then discarding all samples except the highest-scoring one. This approach is mostly used in the black-box setting to watermark LLMs without access to the logits (Bahri & Wieting, 2026), for semantic-aware watermarking (Dabiriaghdam & Wang, 2025; Hou et al., 2024; Huo et al., 2026; Zhang et al., 2025), or to obtain cryptographic guarantees such as undetectability or public detectability (Christ & Gunn, 2024; Fairoze et al., 2024). Importantly, rejection-sampling watermarks are prohibitively costly, as they require sampling 10 tokens for every token generated (Huo et al., 2026).

**Open-Weights Watermarks** Open-weights watermarking works by embedding the watermarking algorithm directly into the model weights. It can be learned (Gu et al., 2024), or embedded into the weights more directly (Block et al., 2025). While this approach could technically be used for closed-source models, open-weights models are not distortion-free, and modifying the weights can lead to unforeseen changes in quality.

**Societal implications of watermarking** All the watermarking methods described are designed to track the provenance of AI-generated text. By relying on statistical testing, they provide guarantees on the expected false-positive rate, thus reducing the risk of false accusation. However, we argue that in high-stakes scenarios watermarks should be used in conjunction with other evidence of wrongdoing.

At the same time, we acknowledge potential misuses of watermarks (e.g. tracking user content without their consent), but ultimately argue that the benefits of watermarking outweigh the risks.

## B Extended Discussion

In this section, we discuss several aspects of our framework in more depth: which insights it provides beyond the expected existence of trade-offs, why we fix the detector, the practical implications of the reduced diversity of soft-constraint schemes, and how the framework could be extended to multi-bit watermarking.

**Insights Provided by the Framework** That watermarks trade off detectability against quality is widely expected. Our framework, however, identifies the fundamental *mechanism* behind the quality-diversity-power trade-off. While prior works have empirically observed that some watermarks may degrade diversity (Gloaguen et al., 2025b; Wu et al., 2026), Theorem F.5 shows that this is not an artifact of specific designs: any optimal watermark with only soft constraints (including all optimal distortion-free watermarks) is necessarily deterministic given the scores. Hence, with a sum-based detector, maximizing power under a distribution-level constraint unavoidably sacrifices diversity, and preserving diversity requires a hard constraint, which in turn costs quality. A second, perhaps surprising, finding is that several popular schemes that were derived through seemingly unrelated bottom-up designs are recovered *exactly* as solutions to our optimization problem. This reveals which constraint each scheme implicitly optimizes (e.g., Red-Green watermarks are optimal with respect to the KL constraint), which we confirm empirically in Sec. 5.1. Note that this experiment is not circular: our theory *predicts* which scheme should be optimal for a given constraint, and the experiment tests this prediction on real LLM outputs, where the constraint is measured over whole sequences and approximations (e.g., the penalized formulation) are used.

**Jointly Optimizing the Detector** Our framework fixes the hashing mechanism and the detector, and optimizes the sampling mechanism. A natural question is whether the detector and the sampling mechanism could be optimized jointly. He et al. (2025) study this joint problem and find that the resulting optimal scheme is not robust: modifying a few tokens suffices to remove the watermark. Hence, jointly optimizing the detector and the generator is not straightforward, and would require explicitly incorporating robustness into the optimization problem. By contrast, with a fixed sum-based detector, robustness naturally arises, as each token contributes (almost) independently to the detection statistic. This is why we opted for optimizing the sampling mechanism only, and we leave the joint problem with robustness constraints to future work.

**Implications of Low Diversity** The low diversity of soft-constraint schemes (Sec. 5.4) means that, under repeated queries with the same prompt, the watermarked model tends to produce similar answers, akin to fixing the sampling seed. This effect is shared with other distortion-free schemes, such as AAR/KTH (Aaronson, 2023; Kuditipudi et al., 2024) and SynthID (Dathathri et al., 2024), the latter being actively deployed. In particular, it can make the presence of a watermark detectable through repeated queries (Gloaguen et al., 2025b). If these are concerns for a given deployment, our framework directly suggests the remedy: using a hard constraint (e.g., the hard-PPL scheme), which retains diversity at the cost of quality.

## C Experimental Details

In this section, we detail the hyperparameters used for our experiments in Sec. 5.

**Watermark Hyperparameters** For Red-Green watermarks, we used as a score distribution a Bernoulli distribution with parameter $\gamma = 0.5$. For AAR/KTH, we used a Gumbel distribution with parameters $(0, 1)$. For SynthID, we optimize the number of layers and the number of candidates per tournament jointly to find the Pareto-optimal parameters. Lastly, for the $\chi^2$ watermark, the hard PPL watermark, and the soft PPL watermark, we used a binomial distribution with parameters $(30, 0.5)$. For the hashing mechanism, we use for all schemes the SumHash hashing scheme from Kirchenbauer et al. (2023) with the four previous tokens as the context. For detection, for Red-Green watermarks we used a binomial test, for AAR/KTH a Kolmogorov–Smirnov test between the observed score

distribution and the Gumbel CDF, for SynthID we used a binomial test, and lastly for $\chi^2$ watermarks and both PPL watermarks, we also used a binomial test. For all tests, following the guidelines from Fernandez et al. (2023), we removed all (context, token) duplicates prior to the statistical test.

**Sampling Hyperparameters** For generating replies, we used the instruct version of LLAMA3.1-8B with a temperature of 0.7 and 16-bit quantization. We also used top-$k$ with $k = 50$ and constrained the model to generate 200-token-long replies. For measuring perplexity, we use the instruct version of QWEN3-30B with a temperature of 1.0 and 8-bit quantization.

**Resources** For inference, we use vLLM (Kwon et al., 2023) on a single NVIDIA RTX PRO 6000 Blackwell GPU.

## D ALGORITHMS

In this section, we examine the different sampling mechanisms for the watermarks introduced in Sec. 4. We recall that, as explained in Sec. 3, we use $\varepsilon$ as a scheme parameter when it corresponds to the constrained formulation (Equation (2)) and $\delta$ when it corresponds to the penalized formulation (Equation (3)).

### D.1 RED-GREEN WATERMARKS

**Algorithm 1** Red-Green Watermark Sampling Mechanism

**input** Next-token probability $p$, watermark strength $\delta$, current scores $g \in \mathbb{R}^{|\Sigma|}$.

$q \leftarrow p \exp(\delta g)$
$q \leftarrow \frac{1}{||q||_1} q$ ▷ *Normalization*
**return:** $q$

### D.2 AAR/KTH WATERMARKS

**Algorithm 2** AAR/KTH Watermark Sampling Mechanism

**input** Next-token probability $p$, watermark strength $\delta$, current scores $g \in \mathbb{R}^{|\Sigma|}$.

$i \leftarrow \arg\max(g + \frac{1}{1+\delta} \log p)$
$q_i \leftarrow 1$
$\forall u \neq i,\ q_u \leftarrow 0$
**return:** $q$

### D.3 CHI2 WATERMARK

**Algorithm 3** Compute Optimal Shift $\mu$

**input** Next-token probability $p$, watermark strength $\delta$, current scores $g \in \mathbb{R}^{|\Sigma|}$.

Sort indices $\pi$ such that $g_{\pi_1} \geq g_{\pi_2} \geq \cdots \geq g_{\pi_{|\Sigma|}}$
Initialize active support $\mathcal{S} \leftarrow \emptyset$
**for** $i = 1$ **to** $|\Sigma|$ **do**
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{\pi_i\}$
  $\mu \leftarrow \frac{1 - \sum_{j \in \mathcal{S}} p_j (1 + \delta g_j)}{\delta \sum_{j \in \mathcal{S}} p_j}$
  **if** $\delta \mu \geq -1 - \delta g_{\pi_i}$ **then**
    **return:** $\mu$
  **end if**
**end for**
**return:** $\mu$

**Algorithm 4** $\chi^2$ Watermark Sampling Mechanism

**input** Next-token probability $p$, watermark strength $\delta$, current scores $g \in \mathbb{R}^{|\Sigma|}$.
1: $\mu \leftarrow \text{Compute_mu}(p, \delta, g)$
2: $q \leftarrow p[1 + \delta(g + \mu)]_+$
3: **return:** $q$

### D.4 PPL WATERMARKS

**Algorithm 5** PPL (Hard) Watermark Sampling Mechanism

**input** Next-token probability $p$, watermark strength $\varepsilon$, current scores $g \in \mathbb{R}^{|\Sigma|}$.
1: $q \leftarrow \text{Solve_LP}(\arg\max_{x \in \Delta(\Sigma)} x \cdot g, \text{ s.t. } -x \cdot \log p \leq -p \cdot \log p + \varepsilon)$ ▷*Use convex-hull solver*
2: **return:** $q$

**Algorithm 6** PPL (Soft) Constraint

**Require:** Next-token probability $p$, watermark strength $\varepsilon$, current scores $g \in \mathbb{R}^{|\Sigma|}$.
**input** $\beta$
1: $i \leftarrow \arg\max(g + \beta \log p)$
2: **return:** $\log p_i - p \cdot \log p + \varepsilon$

**Algorithm 7** PPL (Soft) Watermark Sampling Mechanism

**Require:** Number of Monte Carlo samples $n$.
**input** Next-token probability $p$, watermark strength $\varepsilon$, current scores $g \in \mathbb{R}^{|\Sigma|}$, distribution $G$.
1: $g_1, \ldots, g_n \leftarrow \text{Sample}(G)$ ▷*We sample once at the beginning*
2: $\beta \leftarrow \text{Find_root}(\sum_{i=1}^{n} \text{PPL_Constraint}(p, \varepsilon, g_i))$ ▷*We use a bisection solver.*
3: $i \leftarrow \arg\max(g + \beta \log p)$
4: $q_i \leftarrow 1$
5: $\forall u \neq i,\ q_u \leftarrow 0$
6: **return:** $q$

## E ADDITIONAL EVALUATION

In this section, we include additional ablation experiments to complement our main experiments from Sec. 5. In App. E.1, we extend Sec. 5.1 to the soft KL constraint and the hard PPL constraint. In App. E.2, we ablate the effect of the score distribution $G$ on each of the watermarks from Sec. 4. In App. E.3, we confirm the results from Sec. 5.2 using three additional models (MINISTRAL-3-14B, GEMMA4-12B, and GPT-OSS-20B) and two additional prompt datasets (C4 and OpenMathInstruct-2). In App. E.4, we show the quality–detectability trade-off of each of the watermarks from Sec. 4 using LLM benchmark accuracy as a proxy for quality. Lastly, we provide additional robustness results in App. E.5 and ROC curves in App. E.6.

### E.1 ABLATION ON MEASURING OPTIMALITY WITH RESPECT TO THE CONSTRAINT

Here, we extend the experiments from Sec. 5.1 to include the missing constraints, namely the hard PPL constraint $D_{\text{Hard PPL}}$ and the soft KL constraint $D_{\text{AAR}}$ from Sec. 4.

**Experimental Setup** We follow the same experimental setup as in Sec. 5.1. In particular, for each scheme, we measure the constraint between the original and watermarked next-token probability distributions, and we average it over the sequences. This means that, given $n$ sequences of length $L$ and the corresponding score matrices $G^{(1)}, \ldots, G^{(n)} \in \mathbb{R}^{L \times |\Sigma|}$ and LLM probability distributions

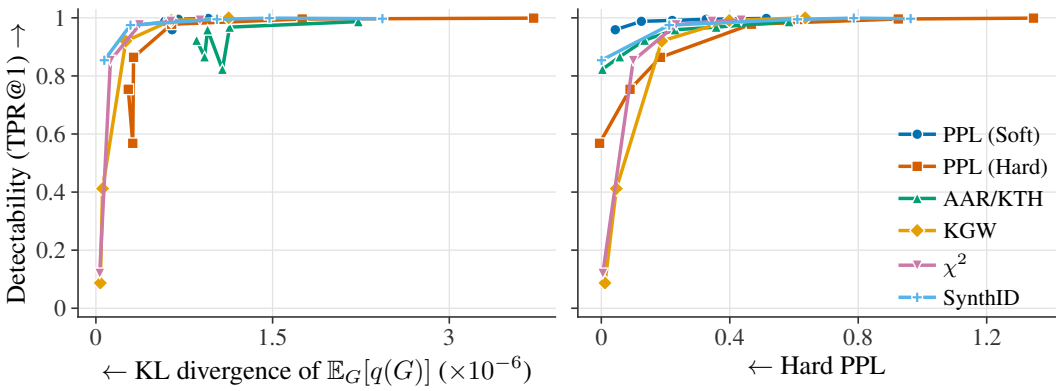



Figure 5: **Additional Detectability–Constraint Trade-off** We compare the trade-off between watermark detectability (TPR@1) and different constraints (KL divergence between $\mathbb{E}_G[q(G)]$ and $p$ (*left*), and hard PPL (*right*)). Responses are 200-token-long replies by LLAMA3.1-8B with temperature 0.7 and 1000 prompts from ELI5.

$p^{(1)}, \ldots, p^{(n)} \in \Delta(\Sigma)^L$, we actually measure

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{L}\sum_{t=1}^{L} D\Big(q(p_t^{(i)}, G_t^{(i)}),\, \mathbb{E}_{G'}\left[q(p_t^{(i)}, G')\right],\, p_t^{(i)}\Big) \approx \mathbb{E}_{p,G}\left[D\Big(q(p, G),\, \mathbb{E}_{G'}\left[q(p, G')\right],\, p\Big)\right]$$

where we explicitly show the dependence of $q$ on $p$. For the hard PPL metric in particular, because it is linear with respect to $q$, this means that in practice we measure the soft PPL metric. A second limitation of our comparison is that not all schemes use the same distribution for $G$. This means that the objective and the detector are not the same; it also means that there is no guarantee that a scheme optimized for a given distribution of $G$ outperforms (in terms of the constraint-quality trade-off) another scheme that is not derived from this particular constraint but uses a different distribution for $G$.

**Results** Figure 5 illustrates both of these limitations. For the KL divergence between the expected watermark distribution and the next-token distribution, we see that AAR, despite being derived from this constraint, is not optimal. This is because, unlike most other schemes, AAR uses the Gumbel distribution for $G$, while other schemes use either the binomial distribution or the Bernoulli distribution. For the hard PPL scheme, similarly, we find that it does not appear to be optimal in Figure 5. This is because, in practice, we can only measure the expected constraint. For the PPL constraint, the expected hard PPL constraint corresponds exactly to the soft PPL constraint, hence we find that Figure 5 (right) and Figure 2 (right) are very similar.

### E.2 Ablation on the Score Distribution

Here, we ablate the score distribution $G$. For all the schemes from Sec. 4 except AAR/KTH and SynthID, there is no fundamental restriction on which distribution to choose for $G$. Yet, our framework does not indicate which score distribution to choose to maximize power: in the optimization problem from Equation (2), the distribution is treated as a constant. This means that we have to find experimentally which distribution is best to maximize the detectability–constraint trade-off for each watermark.

**Experimental Setup** We follow closely the experimental setup from Sec. 5.1. Specifically, for each scheme, we measure its respective constraint between the original and watermarked next-token probability distributions, and we average it over the sequence. We also measure the perplexity (computed with QWEN3-30B) to evaluate the detectability–quality trade-off. Each scheme is evaluated with 6 different distributions for the token scores $G$: Bernoulli with success 0.5, binomial with parameters $(30, 0.5)$, Gumbel with parameters $(0, 1)$, lognormal with parameters $(0, 1)$, normal with parameters $(0, 1)$, and uniform over $[0, 1]$.

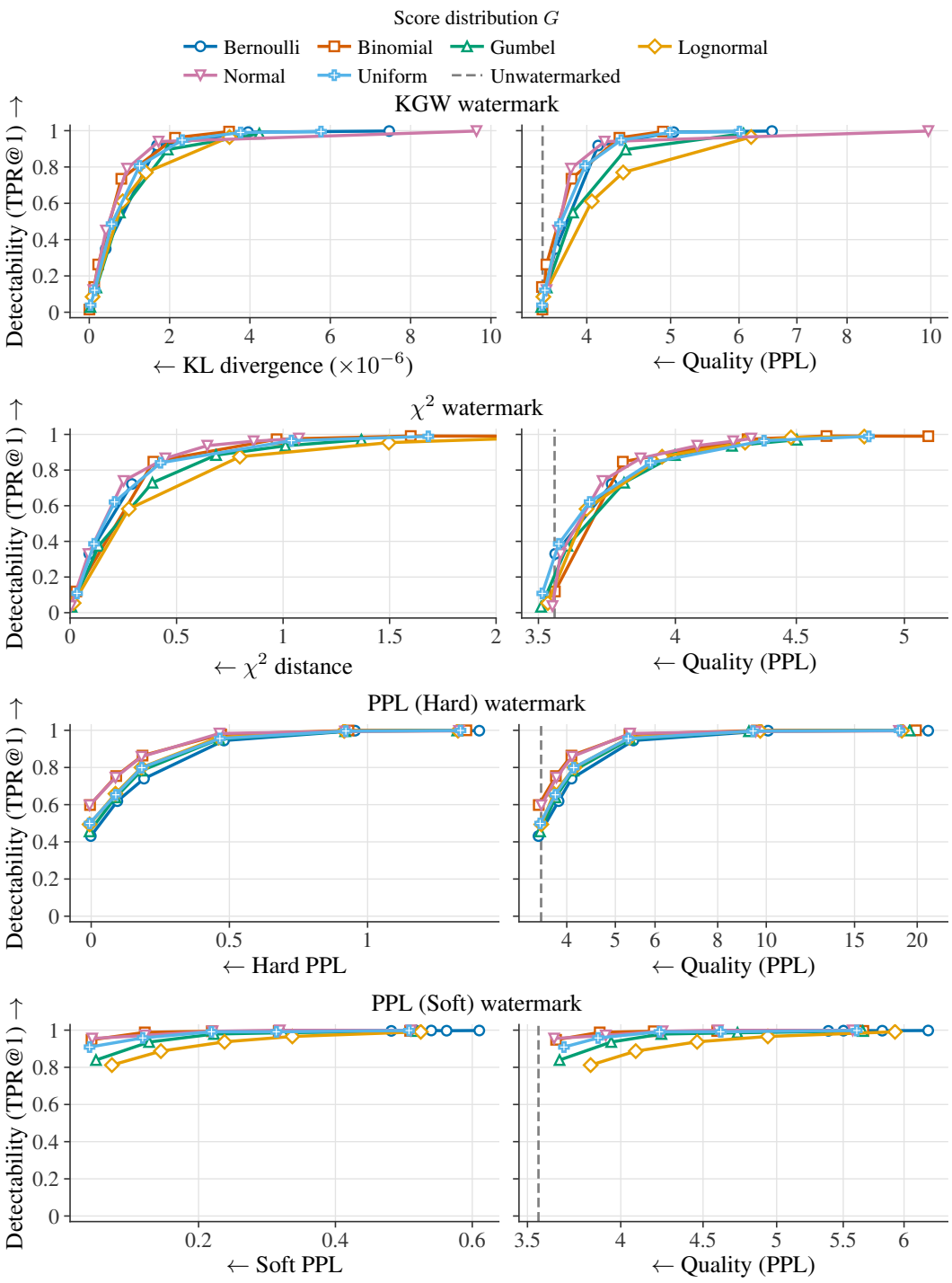



Figure 6: **Watermarks' Trade-offs for Different Score Distributions.** For each watermark, we compare the trade-off between watermark detectability (TPR@1) and its respective constraint (*left*), and between watermark detectability and perplexity (*right*) for different score distributions $G$. The gray dashed line corresponds to the unwatermarked perplexity. Responses are 200-token-long replies by LLAMA3.1-8B with temperature 0.7 and 1000 prompts from ELI5.

**Results** Figure 6 shows for each scheme the different trade-offs when using various score distributions. In particular, we find that for hard-constraint watermarks, the choice of the score distribution

Table 3: **Matched Score Distribution and Detector.** Summary of Figure 6: for each score distribution $G$ (shared by all schemes, together with the corresponding detector), we report the highest TPR@1 whose perplexity is at most the stated budget. "–" means that no evaluated watermark parameter satisfies the budget.

| Distribution | PPL budget | KGW | $\chi^2$ | PPL (Hard) | PPL (Soft) |
|---|---|---|---|---|---|
| Bernoulli | $\leq 4$ | 0.349 | 0.722 | 0.619 | – |
| | $\leq 4.5$ | 0.918 | 0.722 | 0.740 | – |
| | $\leq 5$ | 0.918 | 0.722 | 0.740 | – |
| | $\leq 6$ | 0.992 | 0.722 | 0.945 | 0.998 |
| Binomial | $\leq 4$ | 0.735 | 0.847 | 0.754 | 0.988 |
| | $\leq 4.5$ | 0.961 | 0.973 | 0.864 | 0.994 |
| | $\leq 5$ | 0.995 | 0.991 | 0.864 | 0.995 |
| | $\leq 6$ | 0.995 | 0.991 | 0.978 | 0.996 |
| Gumbel | $\leq 4$ | 0.549 | 0.884 | 0.639 | 0.936 |
| | $\leq 4.5$ | 0.896 | 0.936 | 0.783 | 0.979 |
| | $\leq 5$ | 0.896 | 0.970 | 0.783 | 0.986 |
| | $\leq 6$ | 0.896 | 0.970 | 0.953 | 0.995 |
| Lognormal | $\leq 4$ | 0.085 | 0.875 | 0.659 | 0.812 |
| | $\leq 4.5$ | 0.770 | 0.983 | 0.800 | 0.937 |
| | $\leq 5$ | 0.770 | 0.988 | 0.800 | 0.966 |
| | $\leq 6$ | 0.770 | 0.988 | 0.961 | 0.991 |
| Normal | $\leq 4$ | 0.791 | 0.866 | 0.746 | 0.972 |
| | $\leq 4.5$ | 0.940 | 0.977 | 0.858 | 0.994 |
| | $\leq 5$ | 0.940 | 0.977 | 0.858 | 0.999 |
| | $\leq 6$ | 0.940 | 0.977 | 0.983 | 0.999 |
| Uniform | $\leq 4$ | 0.806 | 0.841 | 0.653 | 0.958 |
| | $\leq 4.5$ | 0.948 | 0.965 | 0.799 | 0.991 |
| | $\leq 5$ | 0.990 | 0.989 | 0.799 | 0.991 |
| | $\leq 6$ | 0.990 | 0.989 | 0.956 | 0.998 |

has almost no influence. For all the distributions tested, the resulting watermarks had similar trade-offs. One exception is the $\chi^2$ watermark with a Bernoulli distribution, where we find that increasing the distortion beyond a threshold leads to catastrophic failures of the watermarking scheme. This is why we only show low distortions for the $\chi^2$ watermark with a Bernoulli distribution in Figure 6.

For soft-constraint watermarks, we evaluate the soft PPL scheme in Figure 6. We find that, in this case, the choice of the distribution $G$ has a significant influence on the quality–detectability trade-off. Specifically, the heavier the tail of the distribution, the lower the performance of the scheme. Interestingly, this is the opposite behavior from what Tsur et al. (2025) observed when extending the AAR scheme to non-Gumbel distributions. This suggests that there is no single best distribution that works for all watermarking schemes.

**Controlled Comparison** Because hard-constraint schemes are optimal for any score distribution, Figure 6 also provides a controlled comparison of the constraints: for a given distribution $G$, all schemes share the same scores and the same detector, and only differ in their constraint. We summarize this comparison in Table 3 at four perplexity budgets. We find that, for all light-tailed distributions (binomial, Gumbel, normal, and uniform), the soft-PPL scheme achieves the best (or jointly best) detectability at every perplexity budget. Hence, the trade-offs observed in Sec. 5 are driven by the constraint rather than by differences in score distributions or detectors. The two exceptions are the heavy-tailed lognormal distribution, as discussed above, and the Bernoulli distribution, for which Algorithm 7 does not apply (see below).

Figure 6 also highlights an interesting result for the soft PPL watermark. As stated in Theorem F.9, Algorithm 7 only holds for continuous distributions (otherwise, to enforce the constraint in addition to $\beta$ one must find the tie-breaking rule). Hence, we see that using Algorithm 7 with a Bernoulli distribution leads to a more distortionary watermark than expected.

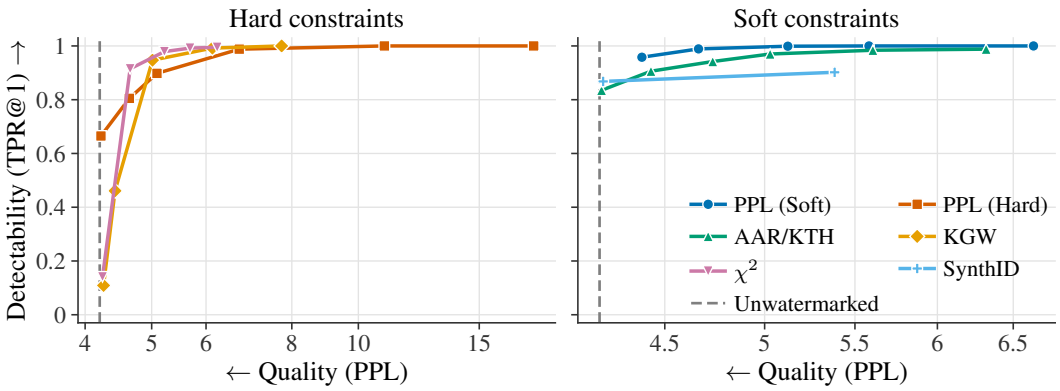



Figure 7: **Comparison of the Detectability–Quality Trade-off.** We compare the trade-off between watermark detectability (TPR@1) and text quality (log PPL) for different constraint instantiations and for different $\varepsilon$. The *left* figure shows the hard constraints and the *right* one shows the soft constraints. The dashed line corresponds to the log PPL of the unwatermarked replies. Responses are 200-token-long replies by MINISTRAL-3-14B with temperature 0.7 and 1000 prompts from ELI5.

Table 4: **Detectability-Quality Trade-off Across Models and Datasets.** Summary of Figure 8: each cell reports the best TPR@1 (%) attained under the stated mean-PPL budget. We **bold** the highest value in each row; "–" means that the scheme was not evaluated in that setting.

| Setting | PPL budget | AAR/KTH | $\chi^2$ | KGW | PPL (Hard) | PPL (Soft) | SynthID |
|---|---|---|---|---|---|---|---|
| GEMMA4-12B · ELI5 | $\leq 6.2$ | 61.9 | 41.2 | 14.2 | 13.9 | 60.3 | **62.9** |
| | $\leq 6.5$ | 68.6 | 75.4 | 36.0 | 24.6 | **76.8** | 62.9 |
| | $\leq 7$ | 68.6 | 75.4 | 66.2 | 38.6 | **95.1** | 87.7 |
| GPT-OSS-20B · ELI5 | $\leq 11$ | 99.8 | 43.4 | – | 99.3 | **99.9** | – |
| | $\leq 12$ | 99.8 | 43.4 | 88.4 | 99.6 | **99.9** | – |
| | $\leq 14$ | 99.9 | 99.8 | 99.8 | 99.6 | **100.0** | – |
| LLAMA3.1-8B · C4 | $\leq 8$ | 99.1 | 49.6 | 57.1 | 99.3 | **99.7** | 99.2 |
| | $\leq 10$ | 99.6 | 99.0 | 98.5 | 99.7 | **99.9** | 99.2 |
| | $\leq 12$ | 99.6 | 99.7 | 99.8 | 99.9 | **100.0** | 99.8 |
| LLAMA3.1-8B · OpenMathInstruct-2 | $\leq 2.5$ | 51.4 | 7.1 | 16.7 | 30.7 | **68.5** | 45.9 |
| | $\leq 3$ | 66.9 | 73.9 | 27.9 | 57.9 | **78.4** | 71.7 |
| | $\leq 4$ | 72.7 | 82.8 | 56.7 | 57.9 | **95.2** | 71.7 |

### E.3 EVALUATION ON ADDITIONAL MODELS AND DATASETS

In this section, we reproduce the experiments from Sec. 5.2 using MINISTRAL-3-14B, GEMMA4-12B, and GPT-OSS-20B instead of LLAMA3.1-8B to ensure that the newly proposed watermarks generalize to models of different sizes and from different families. We also evaluate LLAMA3.1-8B on two additional prompt datasets: the `realnewslike` subset of C4 (Raffel et al., 2020), a completion task with high-entropy continuations, and OpenMathInstruct-2 (Toshniwal et al., 2025), a math instruction dataset with low-entropy responses, where there is little room to alter the next-token distribution.

**Results** Figure 7 shows the detectability-quality trade-off for all watermarks from Sec. 4, with hard-constraint watermarks on the left and soft-constraint watermarks on the right. We find that, for all watermarks except SynthID, the results are very similar to Figure 3. In particular, this means that our framework is model-agnostic and can be applied to most LLMs. Regarding SynthID, we find that at higher distortion levels (not represented in Figure 7), model quality significantly deteriorates.

Figure 8 and Table 4 show the results for the remaining models and datasets. Across all settings, the soft-PPL scheme achieves the best detectability at almost every perplexity budget, confirming that the conclusions from Sec. 5.2 are not specific to LLAMA3.1-8B or to ELI5. On OpenMathInstruct-2, all watermarks become less detectable due to the low entropy of the prompts, and on C4 all watermarks become more detectable due to the higher entropy of completions, yet in both cases the ranking

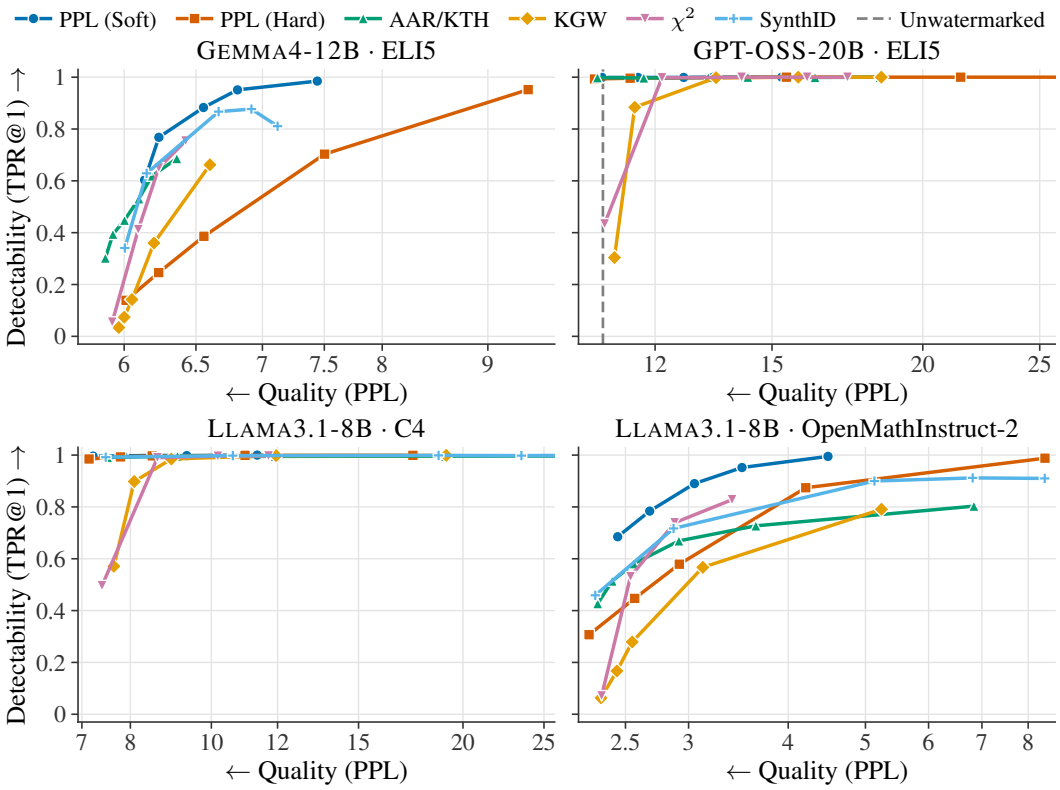



Figure 8: **Detectability–Quality Trade-off Across Models and Datasets.** We compare the trade-off between watermark detectability (TPR@1) and text quality (PPL) for GEMMA4-12B and GPT-OSS-20B on ELI5 (instruction task), and for LLAMA3.1-8B on C4 (completion task) and OpenMathInstruct-2 (low-entropy prompts). For GPT-OSS-20B, the dashed line corresponds to the PPL of the unwatermarked replies. Responses are 200-token-long, with 1000 prompts per setting.

between the schemes remains the same. Notably, the advantage of the soft-PPL scheme is largest in the low-entropy setting (e.g., 68.5% TPR@1 versus 51.4% for AAR/KTH at a PPL budget of 2.5). For GPT-OSS-20B, the high entropy of its outputs makes all schemes except KGW and $\chi^2$ reach near-perfect detectability even without distortion.

### E.4 MEASURING DETECTABILITY-QUALITY TRADE-OFF WITH LLM BENCHMARKS

In this section, we extend the experiments from Sec. 5.2 by measuring the detectability-quality trade-off with LLM benchmark accuracy rather than perplexity. We find that, even when quality is measured through benchmark accuracy, our soft-PPL schemes outperform all prior work.

**Experimental Setup** To measure TPR@1, we follow the experimental setup from Sec. 5.2. For LLM benchmarks, we measure accuracy on three benchmarks with long answers, namely two coding tasks, HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021), and one math task, GSM8K (Cobbe et al., 2021). We use the LM Evaluation Harness framework from Gao et al. (2024). For generating the benchmark answers, we use a temperature of 0.7 and LLAMA3.1-8B.

**Results** Figure 9 shows the detectability-quality trade-off, using LLM benchmark accuracy as a proxy for quality, for all watermarks from Sec. 4, with hard-constraint watermarks on the left and soft-constraint watermarks on the right. We draw a similar conclusion as in Sec. 5.2: both the scheme extensions derived from our framework as well as the novel schemes outperform prior work. In particular, the soft-PPL scheme maximizes the detectability-quality trade-off across all schemes evaluated. Figure 10 shows the trade-off for each benchmark separately, and Table 5 summarizes it at fixed accuracy floors: at the strictest floors, i.e., in the low-distortion regime, the soft-PPL scheme

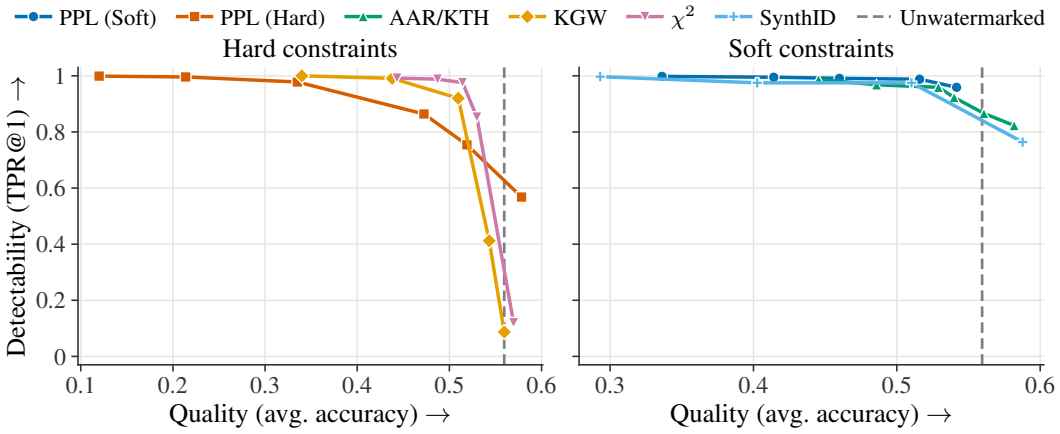



Figure 9: **Comparison of the Detectability-Quality Trade-off.** We compare the trade-off between watermark detectability (TPR@1) and text quality (LLM benchmark accuracy) for different constraint instantiations and for different $\varepsilon$. The *left* figure shows the hard constraints and the *right* one shows the soft constraints. The dashed line corresponds to the average accuracy of the unwatermarked model. Responses are 200-token-long replies by LLAMA3.1-8B with temperature 0.7 and 1000 prompts from ELI5. The benchmark accuracy is averaged over HumanEval, MBPP and GSM8K.

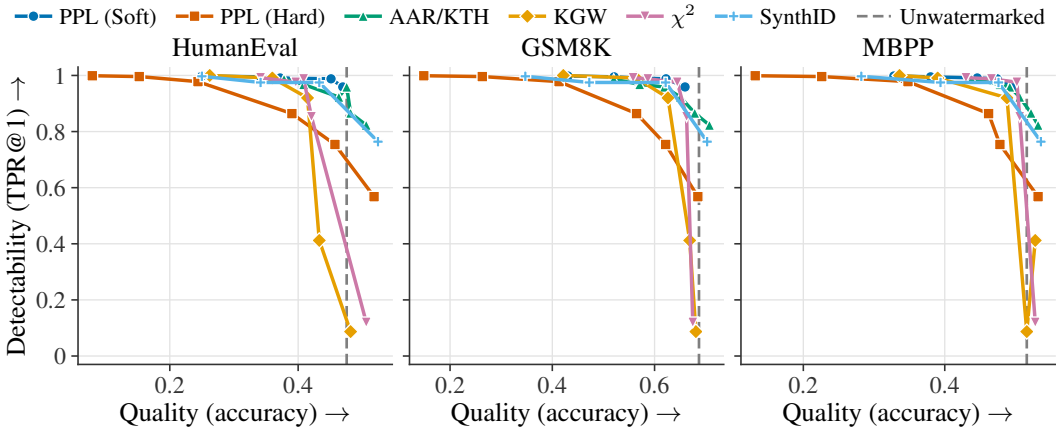



Figure 10: **Detectability-Accuracy Trade-off per Benchmark.** We compare the trade-off between watermark detectability (TPR@1) and accuracy on HumanEval (*left*), GSM8K (*middle*), and MBPP (*right*) for all schemes. The dashed line corresponds to the accuracy of the unwatermarked model.

is always the most (or jointly most) detectable, showing that it preserves model capabilities while maximizing detectability.

### E.5 Evaluating the Robustness of the Proposed Schemes

In this part, we complement the robustness evaluation from Sec. 5.3. We find that, given a fixed impact on quality, the soft-PPL schemes have the highest robustness, and that both hard- and soft-PPL schemes show similar trends to prior work.

**Experimental Setup** We use the same experimental setup as in Sec. 5.1, except that replies are up to 300 tokens long and we keep 1,000 prompts for which every scheme generates at least 200 tokens. We then consider five different edits to the generated text as in prior work (Pan et al., 2024): word deletion, synonym substitution, context-aware substitution (using a BERT model), paraphrasing with GPT-5.4 MINI and with DIPPER (Krishna et al., 2023) (using varying strengths), and back-translation (from English → French → English) using GPT-5.4 MINI. For the watermark hyperparameters, we

Table 5: **Detectability-Accuracy Trade-off.** Summary of Figure 10: each cell reports the best TPR@1 (%) attained while maintaining at least the stated benchmark accuracy. We **bold** the highest value in each row.

| Benchmark | Accuracy | AAR/KTH | $\chi^2$ | KGW | PPL (Hard) | PPL (Soft) | SynthID |
|---|---|---|---|---|---|---|---|
| HumanEval | $\geq 40$ | 96.8 | **98.8** | 92.0 | 75.4 | **98.8** | 97.5 |
| | $\geq 30$ | 98.7 | 99.2 | 99.1 | 86.4 | **99.5** | 97.5 |
| | $\geq 20$ | 98.7 | 99.2 | **100.0** | 97.8 | 99.8 | 99.7 |
| GSM8K | $\geq 60$ | 95.9 | 97.6 | 92.0 | 75.4 | **98.8** | 97.5 |
| | $\geq 50$ | 98.7 | 99.2 | 99.1 | 86.4 | **99.5** | 97.5 |
| | $\geq 40$ | 98.7 | 99.2 | **100.0** | 97.8 | 99.8 | 97.5 |
| MBPP | $\geq 45$ | 96.8 | **98.8** | 92.0 | 86.4 | **98.8** | 97.5 |
| | $\geq 35$ | 98.7 | 99.2 | 99.1 | 86.4 | **99.5** | 97.5 |
| | $\geq 25$ | 98.7 | 99.2 | **100.0** | 97.8 | 99.8 | 99.7 |

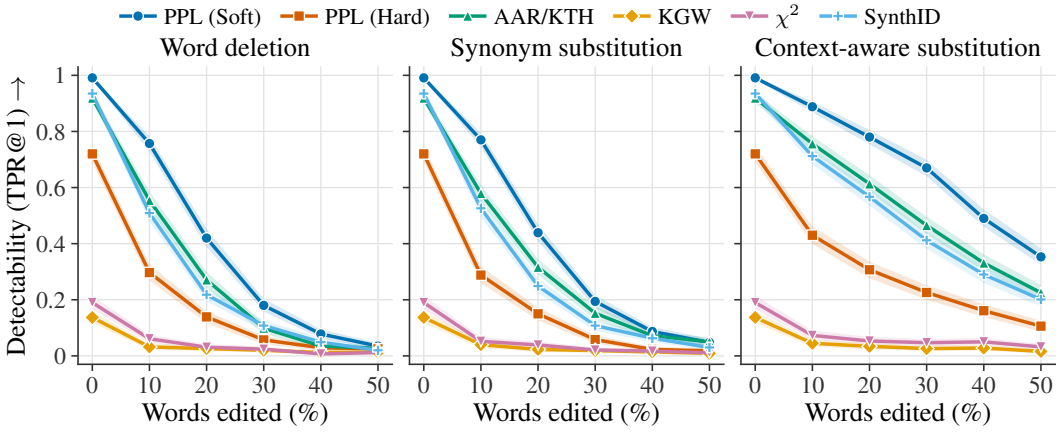



Figure 11: **Robustness against word-editing attacks.** We compare the robustness of watermarking schemes against word-editing attacks. We find that most schemes exhibit similar trends with respect to the percentage of modified words.

choose them such that all schemes are in the low-distortion regime and have the same perplexity of $\approx 3.5$.

**Results** Figure 11 shows the robustness of all considered schemes against word-editing attacks. We find that all schemes have similar trends, and that Soft-PPL outperforms all considered schemes. Table 2 (in Sec. 5.3) shows robustness against back-translation and paraphrasing. For back-translation, Soft-PPL also outperforms all considered schemes. For paraphrasing with GPT-5.4 MINI, all schemes lose most of their signal (TPR@1 of at most 0.13, for Soft-PPL), suggesting that robustness against paraphrasing with a strong paraphraser is still an open issue for token-level watermarks. Indeed, Figure 12 shows robustness against paraphrasing using DIPPER, which allows us to control the paraphrasing strength. For lower strengths, we find that soft-PPL outperforms all considered schemes, but watermark detectability quickly decays for all schemes.

### E.6 EVALUATING THE FPR-TPR TRADE-OFF

In this part, we evaluate the FPR-TPR trade-off (i.e., the ROC curve) of the watermarking schemes from Sec. 5.2. We find that our proposed schemes, soft- and hard-PPL, exhibit similar trends to prior watermarking works, and that they can be used across all practical FPR levels.

**Experimental Setup** We use the same experimental setup as in Sec. 5.1. For the watermark hyperparameters, we choose values such that all schemes are in the low-distortion regime with a perplexity of $\approx 3.5$, and in a higher-distortion regime at $\approx 4.5$.

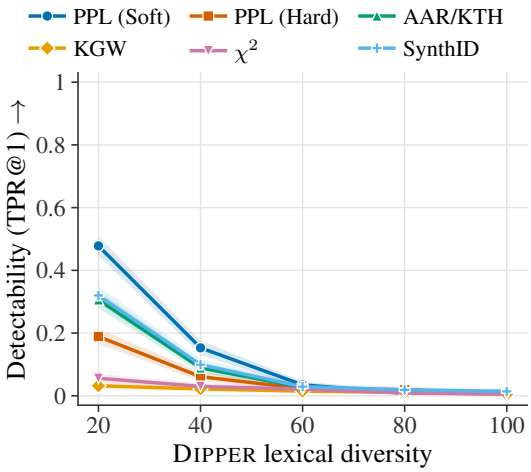



Figure 12: **Robustness against paraphrasing with DIPPER.** We compare the robustness of watermarking schemes against paraphrasing with DIPPER, for different paraphrasing strengths.

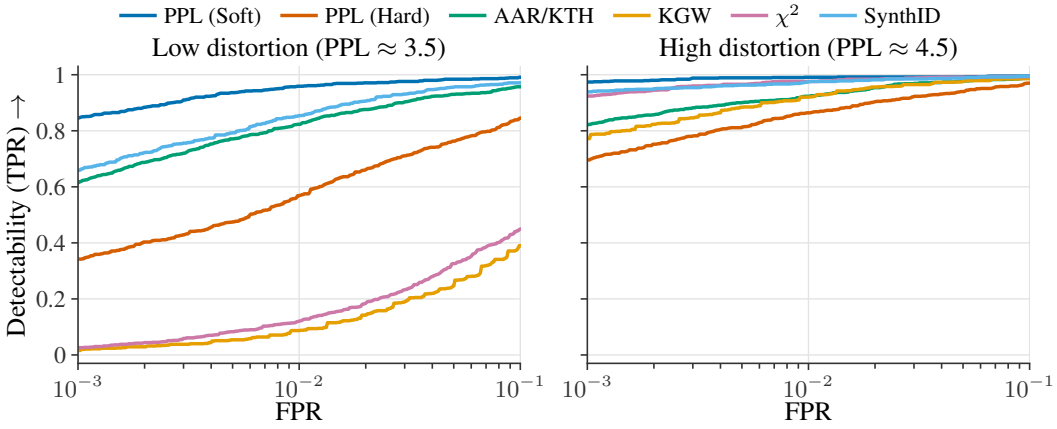



Figure 13: **ROC curves.** We compare the FPR-TPR trade-off of all considered watermarking schemes in the low-distortion regime (left) and high-distortion regime (right). We find that most schemes exhibit similar trends.

**Results** Figure 13 shows that the soft-PPL scheme consistently outperforms prior work at all FPR levels considered. More generally, all schemes exhibit similar trends with respect to the FPR-TPR trade-off.

# F PROOFS

We follow the notation from the main text: $G$ denotes the (context-dependent) score random vector with support $\mathbb{G}$, $g \in \mathbb{G}$ a realization, and $q(g) \in \Delta(\Sigma)$ the corresponding watermarked distribution.

## F.1 RED-GREEN WATERMARKS

Here we formalize and then prove the different results established with respect to Red-Green watermarks. In particular, we show that KL constraints ($D_{\text{R-G}}$) indeed lead to the Red-Green watermark sampling mechanism (Theorem F.1). We also extend our framework with the penalized optimization problem Theorem F.2. The penalized optimization problem is exactly the Red-Green watermark from Kirchenbauer et al. (2023), where the scheme is parametrized by a strength parameter $\delta$. With the constrained version, the parameter $\delta$ is only implicitly defined and depends on $\varepsilon$ and $p$.

**Theorem F.1.** *Let $p \in \Delta(\Sigma)$ have full support, let $g \in \mathbb{G}$ be non-constant, and let $\varepsilon > 0$. Consider*

$$q^*(g) \in \underset{q(g)\in\Delta(\Sigma)}{\arg\max} \quad g \cdot q(g)$$
$$\text{subject to} \quad \mathrm{KL}(q(g)\|p) \le \varepsilon,$$

*where $\mathrm{KL}(q(g)\|p) = \sum_{u\in\Sigma} q(g)_u \log \frac{q(g)_u}{p_u}$ with the convention $0 \log 0 = 0$.*

*Define, for $\delta \ge 0$,*

$$q_\delta(g)_u := \frac{p_u e^{\delta g_u}}{Z_g(\delta)}, \qquad Z_g(\delta) := \sum_{v\in\Sigma} p_v e^{\delta g_v},$$

*and let*

$$M(g) := \arg\max_{u\in\Sigma} g_u, \qquad \varepsilon_{\max}(g) := -\log\Big(\sum_{u\in M(g)} p_u\Big).$$

*Then:*

1. *If $0 < \varepsilon < \varepsilon_{\max}(g)$, there exists a* unique *$\delta \in (0,\infty)$ such that $\mathrm{KL}(q_\delta(g)\|p) = \varepsilon$. Moreover, the optimizer is unique and equals $q^*(g) = q_\delta(g)$, i.e.*
$$q^*(g) \propto p\, e^{\delta g}.$$

2. *If $\varepsilon \ge \varepsilon_{\max}(g)$, the distribution*
$$q_\infty(g)_u := \frac{p_u \mathbf{1}\{u \in M(g)\}}{\sum_{v\in M(g)} p_v}$$
*is optimal.*

*Proof.* Fix $g \in \mathbb{G}$. The problem is a convex optimization problem maximizing a linear function over a convex set (the KL-divergence ball). We introduce a Lagrange multiplier $\lambda \ge 0$ for the inequality constraint $\mathrm{KL}(q(g)\|p) - \varepsilon \le 0$ and a multiplier $\nu \in \mathbb{R}$ for the simplex constraint $\sum_u q(g)_u = 1$. The Lagrangian is:

$$\mathcal{L}(q(g),\lambda,\nu) = -\sum_{u\in\Sigma} q(g)_u g_u + \lambda\left(\sum_{u\in\Sigma} q(g)_u \log\frac{q(g)_u}{p_u} - \varepsilon\right) + \nu\left(\sum_{u\in\Sigma} q(g)_u - 1\right).$$

Note that we minimize $-\mathcal{L}$ to maximize the objective. Taking the derivative with respect to $q(g)_u$ and setting it to zero:

$$\frac{\partial\mathcal{L}}{\partial q(g)_u} = -g_u + \lambda\left(\log\frac{q(g)_u}{p_u} + 1\right) + \nu = 0.$$

Rearranging for $q(g)_u$:

$$\lambda \log\frac{q(g)_u}{p_u} = g_u - \nu - \lambda \implies q(g)_u = p_u \exp\left(\frac{g_u - \nu - \lambda}{\lambda}\right). \tag{13}$$

If $\lambda = 0$, equation Equation (13) implies $g_u = \nu$ for all $u$ where $q(g)_u > 0$, which contradicts the assumption that $g$ is non-constant. Thus, we must have $\lambda > 0$. Let $\delta := 1/\lambda \in (0,\infty)$. Absorbing the normalization constants into a factor $1/Z_g(\delta)$, we obtain the form:

$$q_\delta(g)_u = \frac{p_u e^{\delta g_u}}{Z_g(\delta)}, \quad \text{where } Z_g(\delta) = \sum_{v\in\Sigma} p_v e^{\delta g_v}.$$

By the Karush-Kuhn-Tucker (KKT) conditions, specifically complementary slackness, we require $\lambda(\mathrm{KL}(q_\delta(g)\|p) - \varepsilon) = 0$. Since $\lambda > 0$, we must have $\mathrm{KL}(q_\delta(g)\|p) = \varepsilon$, unless the constraint is inactive at the global maximum of the objective over the simplex (Case 2).

Let $\psi_g(\delta) := \mathrm{KL}(q_\delta(g)\|p)$. We analyze the behavior of $\psi_g(\delta)$:

- As $\delta \to 0$, $q_\delta(g) \to p$, so $\psi_g(0) = 0$.

- As $\delta \to \infty$, the mass of $q_\delta(g)$ concentrates on the set $M(g) = \arg\max_u g_u$. Specifically, $q_\delta(g) \to q_\infty(g)$, where $q_\infty(g)_u = \frac{p_u \mathbf{1}\{u \in M(g)\}}{\sum_{v \in M(g)} p_v}$.
- The derivative is $\psi_g'(\delta) = \delta \operatorname{Var}_{U \sim q_\delta(g)}(g_U)$. Since $g$ is non-constant, the variance is positive, so $\psi_g(\delta)$ is strictly increasing.

The limit of the KL divergence is

$$\begin{aligned}\lim_{\delta \to \infty} \psi_g(\delta) &= \sum_{u \in M(g)} \frac{p_u}{\sum_{v \in M(g)} p_v} \log\left(\frac{p_u / \sum_{v \in M(g)} p_v}{p_u}\right) \\ &= -\log\left(\sum_{v \in M(g)} p_v\right) \\ &= \varepsilon_{\max}(g).\end{aligned}$$

**Case 1:** $0 < \varepsilon < \varepsilon_{\max}(g)$**.**
Since $\psi_g(0) = 0$, $\lim_{\delta \to \infty} \psi_g(\delta) = \varepsilon_{\max}(g)$, and $\psi_g$ is continuous and strictly increasing, there exists a unique $\delta \in (0, \infty)$ such that $\psi_g(\delta) = \varepsilon$. This $\delta$ satisfies the KKT conditions, making $q_\delta(g)$ the unique optimizer.

**Case 2:** $\varepsilon \geq \varepsilon_{\max}(g)$**.**
The unconstrained maximum of the linear function $g \cdot q(g)$ over the probability simplex $\Delta(\Sigma)$ occurs at the vertices corresponding to $M(g)$. Among distributions supported on $M(g)$, $q_\infty(g)$ minimizes the KL divergence to $p$ (it is the projection of $p$ onto the face defined by $M(g)$). Since $\mathrm{KL}(q_\infty(g) \| p) = \varepsilon_{\max}(g) \leq \varepsilon$, $q_\infty(g)$ is feasible. Since it achieves the global maximum of the objective function $g \cdot q(g)$, it is optimal. □

**Theorem F.2.** *Let $p \in \Delta(\Sigma)$ have full support, let $g \in \mathbb{G}$ be non-constant, and let $\delta > 0$. Consider*

$$q^*(g) \in \operatorname*{arg\,max}_{q(g) \in \Delta(\Sigma)} \quad g \cdot q(g) - \frac{1}{\delta} \mathrm{KL}(q(g) \| p),$$

*We then have,*

$$q^*(g) \propto p\, e^{\delta g}.$$

*Proof.* Fix $g \in \mathbb{G}$. We formulate the Lagrangian $\mathcal{L}(q(g), \nu)$ for the optimization problem, introducing a multiplier $\nu$ for the normalization constraint $\sum_u q(g)_u = 1$:

$$\mathcal{L}(q(g), \nu) = \sum_{u \in \Sigma} q(g)_u g_u - \frac{1}{\delta} \sum_{u \in \Sigma} q(g)_u \ln \frac{q(g)_u}{p_u} - \nu \left(\sum_{u \in \Sigma} q(g)_u - 1\right).$$

The first-order optimality condition with respect to $q(g)_u$ is given by setting the derivative to zero:

$$\frac{\partial \mathcal{L}}{\partial q(g)_u} = g_u - \frac{1}{\delta}\left(\ln \frac{q(g)_u}{p_u} + 1\right) - \nu = 0.$$

Rearranging the terms yields:

$$\ln \frac{q(g)_u}{p_u} = \delta g_u - (1 + \delta\nu) \implies q(g)_u = p_u e^{\delta g_u} e^{-(1+\delta\nu)}.$$

Since $e^{-(1+\delta\nu)}$ is a constant scaling factor independent of $u$ (determined by the constraint $\sum_u q(g)_u = 1$), we conclude that

$$q^*(g)_u \propto p_u e^{\delta g_u}.$$

□

### F.2 AAR/KTH WATERMARKS

Here we formalize and then prove the different results established with respect to AAR/KTH watermarks. We first show the shape of the optimal solution when using soft-KL constraints with any score distribution (Theorem F.3), and then establish the extended watermark mechanism of AAR/KTH (Sec. 4.1) when using a Gumbel distribution (Theorems F.3 and F.4). Lastly, we show that for soft-constrained watermarks with a continuous score distribution, the optimal solution is deterministic (Theorem F.5), which theoretically demonstrates the lack of diversity in soft-constrained schemes observed in Sec. 5.4.

**Theorem F.3.** *Let $p \in \Delta(\Sigma)$ have full support, let $G$ be a random vector with support $\mathbb{G} \subseteq \mathbb{R}^{|\Sigma|}$ and i.i.d. coordinates from a law $\mathcal{L}$, and let $\varepsilon \geq 0$. Consider the optimization problem*

$$q^* \in \underset{q:\mathbb{G}\to\Delta(\Sigma)}{\arg\max} \quad \mathbb{E}\big[G \cdot q(G)\big] \tag{14}$$

$$\textit{subject to} \quad D_{\mathrm{KL}}\big(\mathbb{E}[q(G)]\|p\big) \leq \varepsilon, \tag{15}$$

*where $D_{\mathrm{KL}}(r\|p) = \sum_{u\in\Sigma} r_u \log \frac{r_u}{p_u}$.*

*Then there exists a vector $\tau \in \mathbb{R}^{|\Sigma|}$ and a measurable tie-breaking function $T : \mathbb{G} \times \Sigma \to [0,1]$ such that an optimal solution is given by*

$$q^*(G)_u \;=\; \begin{cases} 1 & \text{if } \{u\} = M(G,\tau), \\ T_u(G) & \text{if } u \in M(G,\tau) \text{ and } |M(G,\tau)| > 1, \\ 0 & \text{otherwise}, \end{cases} \tag{16}$$

*where $M(G,\tau)$ denotes the set of maximizing indices:*

$$M(G,\tau) = \underset{v\in\Sigma}{\arg\max}\,\big(G_v + \tau_v\big).$$

*The tie-breaking weights satisfy $\sum_{u\in M(G,\tau)} T_u(G) = 1$ for all $G$, and $T, \tau$ are chosen such that $D_{\mathrm{KL}}(\mathbb{E}[q^*(G)]\|p) = \varepsilon$ (unless $\tau = 0$ is feasible).*

*In particular, if $G$ has i.i.d. standard Gumbel coordinates, there exists $\alpha \in [0,1]$ such that an optimal solution is given by*

$$\forall u \in \Sigma, \qquad q^*(G)_u \;=\; \mathbb{1}\Big\{u \in \underset{v\in\Sigma}{\arg\max}\,\big(G_v + \alpha \log p_v\big)\Big\},$$

*and $\alpha$ is chosen so that $\mathrm{KL}(\mathbb{E}[q^*(G)]\|p) = \varepsilon$ if feasible, else $\alpha = 0$.*

*Proof.* Let $d = |\Sigma|$ and write indices as $u \in \Sigma$. Set

$$r \;=\; \mathbb{E}[q(G)] \in \Delta(\Sigma).$$

We can rewrite Equation (14) as

$$\max_{q,r} \quad \mathbb{E}\big[G \cdot q(G)\big] \tag{17}$$

$$\text{s.t.} \quad \mathbb{E}[q(G)] = r, \tag{18}$$

$$\mathrm{KL}(r\|p) \leq \varepsilon, \tag{19}$$

$$q(G) \in \Delta(\Sigma) \quad \text{for almost every } G. \tag{20}$$

The feasible set is convex and admits a Slater point, namely $q(G) \equiv p$ (so $r = p$ and $\mathrm{KL}(p\|p) = 0 < \varepsilon$), hence strong duality and KKT conditions hold.

**Shape of an optimizer.** Introduce $\lambda \in \mathbb{R}^d$ as dual variable for Equation (18) and $\mu \geq 0$ for Equation (19). The Lagrangian is

$$\mathcal{L}(q,r,\lambda,\mu) = \mathbb{E}\big[G \cdot q(G)\big] + \lambda \cdot (\mathbb{E}[q(G)] - r) - \mu\big(\mathrm{KL}(r\|p) - \varepsilon\big),$$

which we rewrite as

$$\mathcal{L}(q,r,\lambda,\mu) = \mathbb{E}\big[(G+\lambda) \cdot q(G)\big] - \lambda \cdot r - \mu\big(\mathrm{KL}(r\|p) - \varepsilon\big). \tag{21}$$

For fixed $(\lambda, \mu, r)$, maximizing Equation (21) over $q$ subject to Equation (20) separates over $G$:

$$\sup_q \mathcal{L}(q, r, \lambda, \mu) = -\lambda \cdot r - \mu\big(\mathrm{KL}(r\|p) - \varepsilon\big) + \mathbb{E}\Big[\sup_{q(G)\in\Delta(\Sigma)} (G+\lambda)\cdot q(G)\Big]. \tag{22}$$

For any $g \in \mathbb{G}$,

$$\sup_{q\in\Delta(\Sigma)} (g+\lambda)\cdot q = \max_{v\in\Sigma}(g_v + \lambda_v),$$

and the supremum is attained by any $q$ supported on the maximizers.

Let $(\lambda^*, \mu^*)$ be an optimal dual pair and define

$$M(G, \lambda^*) \;=\; \arg\max_{v\in\Sigma} \big(G_v + \lambda^*_v\big).$$

By strong duality, there exists a primal optimizer $q^*$ that attains the supremum in Equation (22) for $(\lambda^*, \mu^*)$. Thus, for almost every $G$, $q^*(G)$ is supported on $M(G, \lambda^*)$.

Choose a measurable tie-breaking function $T : \mathbb{G} \times \Sigma \to [0, 1]$ such that

$$T_u(G) = 0 \quad \text{if } u \notin M(G, \lambda^*), \qquad \sum_{u\in M(G,\lambda^*)} T_u(G) = 1,$$

and, whenever $M(G, \lambda^*) = \{u\}$ is a singleton, we set $T_u(G) = 1$. Then, with $\tau = \lambda^*$,

$$q^*(G)_u \;=\; \begin{cases} 1 & \text{if } \{u\} = M(G,\tau), \\ T_u(G) & \text{if } u \in M(G,\tau) \text{ and } |M(G,\tau)| > 1, \\ 0 & \text{otherwise,} \end{cases}$$

which is exactly the form in Equation (16).

Let $q^0(G)_u \propto \mathbb{1}\Big\{u \in \arg\max_{v\in\Sigma} G_v\Big\}$ be a solution of the unconstrained problem, and $r^0 = \mathbb{E}[q^0(G)]$. If $D_{\mathrm{KL}}(r^0\|p) \le \varepsilon$, then $q^0$ is feasible for Equation (14) and, being optimal for the unconstrained problem, is also optimal for the constrained one; we may take $\tau = 0$. Otherwise $D_{\mathrm{KL}}(r^0\|p) > \varepsilon$, so the KL constraint is active at any optimum by KKT/complementary slackness. Thus there exists an optimal pair $(q^*, r^*)$ with $D_{\mathrm{KL}}(r^*\|p) = \varepsilon$, and by the previous paragraph $q^*$ has the form Equation (16) for some $\tau$ and measurable tie-breaking rule $T$. This proves the first part of the theorem.

**Gumbel case.** Assume now that $G$ has i.i.d. standard Gumbel coordinates. For any $\tau \in \mathbb{R}^d$, define

$$q^\tau(G)_u \;=\; \mathbb{1}\Big\{u = \arg\max_{v\in\Sigma} \big(G_v + \tau_v\big)\Big\}, \qquad r(\tau)_u \;=\; \mathbb{E}[q^\tau(G)_u]. \tag{23}$$

The Gumbel-max trick gives

$$r(\tau)_u \;=\; \frac{e^{\tau_u}}{\sum_{v\in\Sigma} e^{\tau_v}}, \qquad u \in \Sigma. \tag{24}$$

Fix $\tau$ and write $Z_v = \tau_v + G_v$. For each realization of $G$,

$$\max_v Z_v = \sum_{u\in\Sigma} Z_u\, q^\tau(G)_u = G \cdot q^\tau(G) + \tau \cdot q^\tau(G).$$

Taking expectations and denoting $r = r(\tau)$,

$$\mathbb{E}\big[G \cdot q^\tau(G)\big] = \mathbb{E}\big[\max_v Z_v\big] - \tau \cdot r. \tag{25}$$

From the computation of the distribution of $Z^* = \max_v Z_v$,

$$\mathbb{E}\big[\max_v Z_v\big] = \gamma + \log\Big(\sum_{v\in\Sigma} e^{\tau_v}\Big), \tag{26}$$

where $\gamma$ is the Euler-Mascheroni constant. Combining Equation (25) and Equation (26),

$$\mathbb{E}\big[G \cdot q^{\tau}(G)\big] = \gamma + \log\Big(\sum_{v} e^{\tau_v}\Big) - \tau \cdot r. \tag{27}$$

Yet, we have

$$\tau \cdot r = \sum_{u} r_u \log r_u + \log\Big(\sum_{v\in\Sigma} e^{\tau_v}\Big). \tag{28}$$

Substituting Equation (28) into Equation (27) yields

$$\mathbb{E}\big[G \cdot q^{\tau}(G)\big] = \gamma - \sum_{u\in\Sigma} r_u \log r_u = \gamma + H(r),$$

where $H(r) := -\sum_u r_u \log r_u$ is the Shannon entropy. Thus, in the Gumbel case, optimizing Equation (14) over $q$ is equivalent to optimizing over $r = \mathbb{E}[q(G)]$:

$$\max_{r\in\Delta(\Sigma)} \quad H(r) \tag{29}$$

$$\text{s.t.} \quad \mathrm{KL}(r\|p) \le \varepsilon. \tag{30}$$

Introduce multipliers $\lambda \ge 0$ for Equation (30) and $\beta \in \mathbb{R}$ for $\sum_u r_u = 1$. The Lagrangian is

$$\mathcal{J}(r,\lambda,\beta) = -\sum_u r_u \log r_u - \lambda\left(\sum_u r_u \log\frac{r_u}{p_u} - \varepsilon\right) + \beta\left(\sum_u r_u - 1\right). \tag{31}$$

Assuming $r_u > 0$, stationarity in $r_u$ gives

$$-(\log r_u + 1) - \lambda\big(\log(r_u/p_u) + 1\big) + \beta = 0, \qquad u \in \Sigma.$$

Thus

$$r_u = e^{C} p_u^{\alpha},$$

where

$$\alpha := \frac{\lambda}{1+\lambda} \in [0,1), \qquad C := \frac{\beta - (1+\lambda)}{1+\lambda}. \tag{32}$$

Normalization $\sum_u r_u = 1$ fixes $e^C$, and we obtain the family

$$r_u^*(\alpha) = \frac{p_u^{\alpha}}{\sum_{v\in\Sigma} p_v^{\alpha}}, \qquad \alpha \in [0,1). \tag{33}$$

Note that $r^*(0)$ is the uniform distribution on $\Sigma$, and $r^*(1) = p$. The map $\alpha \mapsto \mathrm{KL}(r^*(\alpha)\|p)$ is continuous, with

$$\mathrm{KL}\big(r^*(0)\|p\big) = \mathrm{KL}(\mathrm{Unif}\|p), \qquad \mathrm{KL}\big(r^*(1)\|p\big) = 0.$$

Hence:

- If $\varepsilon \ge \mathrm{KL}(\mathrm{Unif}\|p)$, the KL constraint is slack and the unconstrained maximizer of $H(r)$ is the uniform distribution $r^*(0)$, corresponding to $\alpha = 0$.
- If $0 \le \varepsilon \le \mathrm{KL}(\mathrm{Unif}\|p)$, there exists (by continuity) some $\alpha \in [0,1]$ such that

  $$\mathrm{KL}\big(r^*(\alpha)\|p\big) = \varepsilon,$$

  and the constraint is active at the optimum.

For any optimal solution $r^*$ of Equation (29)–Equation (30), the softmax bijection Equation (24) allows us to realize $r^*$ as $r(\tau)$ for some $\tau$. From Equation (33), we have

$$\log r_u^*(\alpha) = \alpha \log p_u + \text{const}, \qquad u \in \Sigma.$$

Hence, we see that $\tau_u$ can be chosen of the form

$$\tau_u = \alpha \log p_u + c, \qquad u \in \Sigma,$$

for some constant $c \in \mathbb{R}$. Adding a constant to all coordinates of $\tau$ does not change the argmax in Equation (23), so without loss of generality we may take $c = 0$, that is,

$$\tau_u = \alpha \log p_u, \qquad u \in \Sigma.$$

Combining this with Equation (23) yields

$$q^*(G)_u = \mathbb{1}\Big\{u = \arg\max_{v \in \Sigma} \big(G_v + \alpha \log p_v\big)\Big\},$$

for some $\alpha \in [0, 1]$ chosen as above. This completes the proof. □

**Theorem F.4.** *Let $p \in \Delta(\Sigma)$ have full support, let $G$ be a random vector with support $\mathbb{G} \subseteq \mathbb{R}^{|\Sigma|}$ and i.i.d. coordinates from a Gumbel distribution, and let $\delta > 0$. Consider*

$$q^* \in \arg\max_{q:\mathbb{G}\to\Delta(\Sigma)} \quad \mathbb{E}\big[G \cdot q(G)\big] - \frac{1}{\delta}\mathrm{KL}\big(\mathbb{E}[q(G)] \| p\big),$$

*where $\mathrm{KL}(r\|p) = \sum_{u\in\Sigma} r_u \log \frac{r_u}{p_u}$. Then, we have*

$$\forall u \in \Sigma, \qquad q^*(G)_u \;=\; \mathbb{1}\Big\{u \in \arg\max_{v \in \Sigma} \big(G_v + \frac{1}{1+\delta} \log p_v\big)\Big\}.$$

*Proof.* The proof is similar to that of Theorem F.3. In Equation (31), instead of being a Lagrangian variable, $\lambda$ is explicitly given as $\lambda = \frac{1}{\delta}$. This then gives, in Equation (32), $\alpha = \frac{1}{1+\delta}$. Thus, we have that,

$$\forall u \in \Sigma, \qquad q^*(G)_u \;=\; \mathbb{1}\Big\{u \in \arg\max_{v \in \Sigma} \big(G_v + \frac{1}{1+\delta} \log p_v\big)\Big\},$$

which completes the proof. □

**Theorem F.5.** *Let $p \in \Delta(\Sigma)$, let $G$ be a random vector with support $\mathbb{G} \subseteq \mathbb{R}^{|\Sigma|}$ and i.i.d. coordinates from a continuous law $\mathcal{L}$ with finite first-order moment, and let $\varepsilon \geq 0$. Consider*

$$\begin{aligned} q^* \in \arg\max_{q:\mathbb{G}\to\Delta(\Sigma)} \quad & \mathbb{E}\big[G \cdot q(G)\big] \\ \textit{subject to} \quad & D\big(\mathbb{E}[q(G)], p\big) \leq \varepsilon, \end{aligned}$$

*where $D : \Delta(\Sigma) \times \Delta(\Sigma) \to \mathbb{R}$. Then, if the problem is feasible, there exists an optimal solution with support of size 1.*

*Proof.* Let $q$ be any optimal solution. Let $X \in \Sigma$ be a random variable corresponding to the sampled token, and define

$$\forall g \in \mathbb{G},\ \forall u \in \Sigma, \quad \mathbb{P}[X = u \mid G = g] = q_u(g).$$

Then the objective rewrites as

$$\mathbb{E}[G \cdot q(G)] = \mathbb{E}[G_X],$$

and the average output distribution is the marginal of $X$,

$$r(q)_u \coloneqq \mathbb{E}[q_u(G)] = \mathbb{P}[X = u], \quad u \in \Sigma.$$

The law of $G$ is atomless since its coordinates are i.i.d. from a continuous law. By the Dvoretsky-Wald-Wolfowitz purification theorem (Dvoretzky et al., 1951) applied to $g \mapsto q(g)$ and the finite family of integrable functions

$$(g, u) \mapsto g_u \quad \text{and} \quad (g, u) \mapsto \mathbf{1}\{u = v\},\ v \in \Sigma,$$

there exists a measurable function $f : \mathbb{G} \to \Sigma$ such that, with $X' = f(G)$,

$$\mathbb{E}[G_{X'}] = \mathbb{E}[G_X] \quad \text{and} \quad \mathbb{P}[X' = u] = \mathbb{P}[X = u],\ \forall u \in \Sigma.$$

Define $q' : \mathbb{G} \to \Delta(\Sigma)$ by $q'(g) = e_{f(g)}$, so that $q'(G)$ has support of size 1 almost surely. Then

$$\mathbb{E}[G \cdot q'(G)] = \mathbb{E}[G_{X'}] = \mathbb{E}[G_X] = \mathbb{E}[G \cdot q(G)],$$

and

$$\mathbb{E}[q'(G)]_u = \mathbb{P}[X' = u] = \mathbb{P}[X = u] = \mathbb{E}[q(G)]_u, \quad u \in \Sigma.$$

Hence $D(\mathbb{E}[q'(G)], p) = D(\mathbb{E}[q(G)], p) \leq \varepsilon$, so $q'$ is feasible and optimal. This proves the existence of an optimal solution with support of size 1. □

### F.3 CHI2 WATERMARKS

Here we formalize and then prove the various results established for the one-layer SynthID and $\chi^2$ watermark. In particular, we first establish the general watermark sampling mechanism for the $\chi^2$ watermark in Theorem F.6, and show in Theorem F.7 how it reduces to the 1-layer SynthID tournament in some specific cases.

**Theorem F.6.** *Let $p \in \Delta(\Sigma)$ have full support, let $g \in \mathbb{G}$, and let $\varepsilon > 0$. Consider*

$$q^*(g) \in \underset{q(g)\in\Delta(\Sigma)}{\arg\max} \quad g \cdot q(g)$$
$$\text{subject to} \quad \sum_{u\in\Sigma} \frac{(q(g)_u - p_u)^2}{p_u} \leq \varepsilon.$$

*Then there exists $\delta, \mu \in \mathbb{R}$ such that,*

$$\forall u \in \Sigma, q^*(g)_u = p_u[1 + \delta(g_u + \mu)]_+$$

*where $[x]_+ := \max(0, x)$, $\mu$ is chosen such that $\sum_{u\in\Sigma} q^*(g)_u = 1$, and $\delta$ is chosen such that $\sum_{u\in\Sigma} \frac{(q^*(g)_u - p_u)^2}{p_u} = \varepsilon$ if feasible, else the greedy solution is feasible.*

*Proof.* We introduce a Lagrange multiplier $\lambda \geq 0$ for the inequality constraint $\sum_{u\in\Sigma} \frac{(q_u - p_u)^2}{p_u} = \varepsilon$, a multiplier $\mu \in \mathbb{R}$ for the simplex constraint $\sum_u q_u = 1$ and $\nu \in \mathbb{R}_+^{|\Sigma|}$ for the positivity constraints. The Lagrangian is:

$$\mathcal{L}(q, \lambda, \mu) = -\sum_{u\in\Sigma} q_u g_u + \lambda \left( \sum_{u\in\Sigma} \frac{(q_u - p_u)^2}{p_u} - \varepsilon \right) - \mu \left( \sum_{u\in\Sigma} q_u - 1 \right) - \sum_{u\in\Sigma} \nu_u q_u.$$

Note that we minimize $-\mathcal{L}$ to maximize the objective. Taking the derivative with respect to $q_u$ and setting it to zero:

$$\frac{\partial \mathcal{L}}{\partial q_u} = -g_u + 2\lambda \frac{(q_u - p_u)}{p_u} - \mu - \nu_u = 0.$$

Rearranging for $q_u$:

$$q_u = p_u \left( 1 + \frac{1}{2\lambda}(g_u + \mu + \nu_u) \right).$$

In particular, we have $\forall u \in \Sigma, -q_u \nu_u = 0$ with $\nu_u \geq 0$, which means that,

$$q_u = p_u \left[ 1 + \frac{1}{2\lambda}(g_u + \mu) \right]_+.$$

By letting $\delta = \frac{1}{2\lambda}$ we find the shape from Theorem F.6. □

**Theorem F.7.** *Let $p \in \Delta(\Sigma)$ have full support, let $g \in \mathbb{G}$, and let $\delta > 0$. Consider*

$$q^*(g) \in \underset{q(g)\in\Delta(\Sigma)}{\arg\max} \quad g \cdot q(g) - \frac{1}{2\delta} \sum_{u\in\Sigma} \frac{(q(g)_u - p_u)^2}{p_u}.$$

*Then there exists $\mu \in \mathbb{R}$ such that,*

$$\forall u \in \Sigma, q^*(g)_u = p_u[1 + \delta(g_u + \mu)]_+$$

*where $\mu$ is chosen such that $\sum_{u\in\Sigma} q^*(g)_u = 1$.*

*In particular, if $\delta \leq \frac{1}{\max_{u\in\Sigma} g_u - \min_{u\in\Sigma} g_u}$, then we have*

$$\forall u \in \Sigma, q^*(g)_u = p_u(1 + \delta(g_u - p \cdot g))$$

*Proof.* The derivation of the general form is identical to the first part of Theorem F.6, where $\delta$ is treated as a fixed parameter rather than a Lagrange multiplier.

To prove the second part, assume that the non-negativity constraints are inactive, i.e., for all $u \in \Sigma$, $1 + \delta(g_u + \mu) \geq 0$. Summing the probability mass constraint $\sum q_u = 1$ implies:

$$\sum_{u \in \Sigma} p_u[1 + \delta(g_u + \mu)] = 1 \implies 1 + \delta(p \cdot g + \mu) = 1.$$

Since $\delta > 0$, this yields $\mu = -p \cdot g$. Substituting this back into the solution gives the candidate form $q^*(g)_u = p_u(1 + \delta(g_u - p \cdot g))$.

We must strictly verify that this candidate satisfies the non-negativity assumption, i.e.,

$$\forall u \in \Sigma, \quad 1 + \delta(g_u - p \cdot g) \geq 0 \iff \delta(p \cdot g - g_u) \leq 1.$$

The tightest constraint occurs when $p \cdot g - g_u$ is maximized, which corresponds to $g_u = \min_v g_v$. Thus, we require $\delta \leq \frac{1}{p \cdot g - \min_v g_v}$.

Observing that $p \cdot g \leq \max_v g_v$, we have:

$$p \cdot g - \min_v g_v \leq \max_v g_v - \min_v g_v.$$

Therefore, the condition $\delta \leq \frac{1}{\max_u g_u - \min_u g_u}$ is sufficient to guarantee feasibility. □

### F.4 SynthID Watermark

In this part, we extend our framework to capture the SynthID watermark. In particular, we show that SynthID with $m$-layer tournament corresponds to chaining $m$ optimization problems with $\chi^2$ constraints.

**Theorem F.8.** *Let $p \in \Delta(\Sigma)$ have full support, let $m \in \mathbb{N}$, let $g^{(1)}, \ldots, g^{(m)} \in \{0,1\}^{|\Sigma|}$. Consider the chained optimization problems over $q^{(1)}, \ldots, q^{(m)} \in \Delta(\Sigma)$,*

$$q^{(i)} \in \arg\max_{q \in \Delta(\Sigma)} \quad g^{(i)} \cdot q - \frac{1}{2} \sum_{u \in \Sigma} \frac{(q_u - q_u^{(i-1)})^2}{q_u^{(i-1)}}.$$

*with $q^{(0)} \coloneqq p$. Then, the family defined by*

$$\forall i \in \{1, \ldots, m\}, \forall u \in \Sigma, q_u^{(i)} = q_u^{(i-1)}(1 + (g_u^{(i)} - q_u^{(i-1)} \cdot g^{(i)}))$$

*is an optimal solution.*

*Proof.* Simply proceed by induction. For each step $i$, Theorem F.7 holds, since we have $\delta = 1 \leq 1/(1-0)$. □

### F.5 PPL Watermarks

Here, in Theorem F.9, we formalize and then prove the watermark sampling mechanism established for the soft PPL watermark. For the hard PPL watermark, the resulting constrained optimization problem is a linear programming problem and thus all results stated in Sec. 4.2 are standard results from the linear programming literature.

**Theorem F.9.** *Let $p \in \Delta(\Sigma)$ have full support, let $G$ be a random vector with support $\mathbb{G} \subseteq \mathbb{R}^{|\Sigma|}$ and i.i.d. coordinates from a continuous law $\mathcal{L}$, and let $\varepsilon \geq 0$. Consider the optimization problem*

$$\begin{aligned} q^* \in \arg\max_{q: \mathbb{G} \to \Delta(\Sigma)} \quad & \mathbb{E}\big[G \cdot q(G)\big] \\ \textit{subject to} \quad & (p - \mathbb{E}_G[q(G)]) \cdot \log p \leq \varepsilon. \end{aligned}$$

*Then, an optimal solution is given by*

$$\forall g \in \mathbb{G}, q^*(g) = \arg\max(g_i + \lambda \log p_i),$$

*where $\lambda$ is chosen such that $(p - \mathbb{E}_G[q(G)]) \cdot \log p = \varepsilon$ if feasible, else $\lambda = 0$.*

*Proof.* By Theorem F.5, there exists an optimal solution $q$ with support of size 1 a.s., i.e. there is a measurable $f : \mathbb{G} \to \Sigma$ such that

$$q(g) = e_{f(g)}, \qquad \forall g \in \mathbb{G}.$$

We now solve the problem over such deterministic $q$. We introduce a Lagrange multiplier $\lambda \geq 0$ for the constraint. The Lagrangian is

$$\mathcal{L}(q, \lambda) = \mathbb{E}[G \cdot q(G)] - \lambda\big((p - \mathbb{E}[q(G)]) \cdot \log p - \varepsilon\big) \tag{34}$$

$$= \mathbb{E}\big[(G + \lambda \log p) \cdot q(G)\big] - \lambda\big(p \cdot \log p - \varepsilon\big). \tag{35}$$

For fixed $\lambda$, the term in Equation (35) is maximized by

$$f_\lambda(g) \in \arg\max_{i \in \Sigma} \big(g_i + \lambda \log p_i\big).$$

If the unconstrained maximizer (corresponding to $\lambda = 0$) already satisfies $(p - \mathbb{E}[q^0(G)]) \cdot \log p \leq \varepsilon$, then $\lambda = 0$. Otherwise the constraint is active at the optimum, so $\lambda > 0$ and

$$(p - \mathbb{E}[e_{f_\lambda(G)}]) \cdot \log p = \varepsilon.$$

This concludes the proof. □

## G WATERMARKED TEXT EXAMPLES

This section presents a sample user prompt together with the corresponding watermarked responses. We also report the associated p-values and perplexity scores, where perplexity is computed using QWEN3-30B. The formatting of the LLM response has been lightly edited for readability. Each response is 200 tokens long.

**User Prompt**

Why can't we just taste candy or Sweets and then spit it out to avoid its unhealthy attributes? What makes us swallow it to get satisfaction?

**AAR,** $p = 7.4e-5$**,** $PPL = 4.8$

You're right that our brains are wired to enjoy the taste of sweets, but there are several reasons why we tend to swallow them, even though we know they're unhealthy. Here are a few possible explanations:

1. **Evolutionary adaptations**: For millions of years, humans relied on sugar-rich foods like fruits and berries for energy and survival. As a result, our brains developed a preference for sweet tastes, which signaled the presence of energy-rich compounds. In many cases, it was beneficial to consume these foods and store the energy for later use. Our brains haven't fully adjusted to the modern environment, where sugary foods are abundant but not necessarily necessary for survival.
2. **Reward processing**: Sugar activates the brain's reward centers, releasing feel-good chemicals like dopamine, endorphins, and opiids. These sensations create a strong psychological association between sweetness and pleasure, reinforcing the behavior of consuming sugary foods. Our brains are wired to seek out experiences that stimulate these reward pathways, even if

**Chi2,** $p = 0.0$**,** $PPL = 5.7$

You're right that it would be simpler if our body automatically spat out foods that aren't beneficial to us, but that's not how our digestive system is designed. Here are a few reasons why:

1. **Evolutionary design:** The human body has an amazing ability to adapt and learn, but it's based on the principle of survival and reproduction. Our digestive system has evolved over millions of years to process a wide range of foods, including those that are high in sugar and other energy-rich nutrients. These foods were once scarce, so our bodies were designed to rapidly absorb their nutrients for energy. This allowed our ancestors to survive and thrive in times of scarcity.
2. **The role of the brain:** Our brain plays a significant role in controlling what we eat and how our body processes it. The pleasure and satisfaction we experience when eating something sweet, like candy, is mediated by a system

**KGW,** $p = 0.0$**,** $PPL = 6.4$

You're right that we can choose to spit out candy or sweets if we know they're unhealthy. However, the reason we often can't help but swallow and enjoy these treats is deeply rooted in human psychology, neuroscience, and biology.
Here are some possible explanations:

1. **Evolutionary history**: Throughout human evolution, sweet tastes were a signal of nutrient-dense foods that were scarce and valuable. Our brains evolved to associate sweet tastes with pleasure and reward, encouraging us to consume them and gain energy. In modern times, we still respond to sweet tastes as a cue for pleasure, but our diets no longer match the evolutionary landscape.
2. **Dopamine release**: When we eat something sweet, our brains release dopamine, a neurotransmitter associated with pleasure, reward, and motivation. This release can be intense and immediate, making it difficult to resist the desire to consume more.
3. **Conditioned response**: We often learn to associate certain foods, including sweets, with pleasure and satisfaction

**PPL (Hard),** $p = 1.1e-9$**,** $PPL = 7.4$

It's indeed intriguing as to why our brains would lead us to consume sweets, knowing their potential harm, and yet, we can't help but crave them. There are several biological, psychological, and historical reasons behind this phenomenon.

**Biological reasons:**

1. **Evolutionary legacy:** In the past, sweets were rare and often associated with high-calorie energy sources, making them valuable for survival. Our brains were adapted to consume these energy-dense foods to store fat for potential famines. This instinctual response has been retained, although no longer necessary in modern times with an abundance of food available.
2. **Dopamine release:** Consuming sweets and other pleasurable foods triggers the release of dopamine, a neurotransmitter associated with pleasure signaling. The brain's reward system is activated due to the combination of sugar, fat, and other nutrients releasing dopamine, leading a craving for more.
3. **Lingering taste after spittle:** Some studies indicate even a small amount

**PPL (Soft),** $p = 1.4e-3$**,** $PPL = 5.4$

You're right that it would be ideal if we could just taste sweets and spit them out, avoiding the negative effects of high sugar and calorie intake. However, the process of eating and swallowing is complex, and it's influenced by both physiological and psychological factors. Here are some reasons why we tend to swallow sweets:

1. **Evolutionary origins:** In the past, eating sweet-tasting foods indicated the presence of energy-rich carbohydrates, which were essential for survival. Our ancestors would consume these foods to sustain themselves, especially during times of scarcity. This evolutionary adaptation has been passed down through generations, making us more likely to associate sweet tastes with energy and satisfaction.
2. **Brain chemistry:** When we eat sweets, they activate the brain's reward centers, releasing feelgood hormones like dopamine. This triggers a sense of pleasure and satisfaction, which can lead to overconsumption. The brain's reward system is designed to reinforce behaviors that provide energy and survival benefits, making us more likely to repeat them

**SynthID,** $p = 0.0$**,** $PPL = 13.9$

You're right that it'd seem logical to just taste and spit out candy or sweets to avoid their potential negative effects on our health. However, our bodies have complex neurological and physiological mechanisms that play a significant role in our eating behavior.

Here are some reasons why we tend to swallow the candy and experience the satisfaction we're talking about:

1. **Evolutionary mechanisms**: Throughout history, our brains were designed to respond quickly to available energy sources, like sugary or fatty foods, as these were essential for survival. In the case of sweets, our brain is programmed to respond strongly to glucose (sugar), releasing the neurotransmitter dopamine in the process. This associative reward network drives us to repeat behaviors associated with pleasure, making us reach out for those sweet treats again.
2. **Familiar texture and taste**: When you put something pleasant in your mouth, the sensory experience is unique in several ways: It satisfies your sweet tooth by releasing dopamine in your brain. Our brains are set up to

## H Full Experimental Results

In this section, we report the full numerical results underlying the figures of this paper. All tables are generated from the same data as the corresponding figures.

**Reading the tables** Unless stated otherwise, each row corresponds to one watermark configuration evaluated on 1,000 replies, and we report the TPR at 1% FPR (TPR@1) and at 0.1% FPR (TPR@0.1) in percent. With 1,000 replies, the 95% bootstrap confidence intervals on the TPR have a half-width of at most 3.3 percentage points. The column $\varepsilon$ gives the watermark strength and $G$ the score distribution (Bernoulli, binomial, or Gumbel, see App. C). For SynthID, $\varepsilon$ is instead the number of candidates per tournament, and the $G$ column gives the number of tournament layers $m$. The *Unwatermarked* rows give the quality of the original model, shown as dashed gray lines in the figures.

**Overview** Table 6 reports the detectability-quality trade-off on LLAMA3.1-8B and ELI5 (Sec. 5.1 and Sec. 5.2), including all the constraints from Figures 2 and 5. Tables 7–9 report the same trade-off on the additional models and datasets of App. E.3. Tables 10 and 11 report the score distribution ablation of App. E.2. Table 12 reports the LLM benchmark accuracies of App. E.4. Table 13 reports the word-edit and DIPPER robustness results of App. E.5 (the back-translation and paraphrasing results are already in Table 2). Table 14 reports the ROC curves of App. E.6 at fixed FPRs. Lastly, Table 15 reports the diversity-quality results of Sec. 5.4.

Table 6: Full results on LLAMA3.1-8B and ELI5 (Figures 2, 3 and 5). We report the TPR at 1% and 0.1% FPR, the mean perplexity, and the per-token mean of each constraint between the watermarked and the original next-token distributions. For SynthID, $\varepsilon$ is the number of candidates per tournament and $m$ the number of tournament layers (with Bernoulli$(0.5)$ g-values).

| | | | Detectability (%) ↑ | | Quality ↓ | Constraint (per-token mean) ↓ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scheme | $\varepsilon$ | $G$ | TPR@1 | TPR@0.1 | PPL | KL | $\chi^2$ | Hard PPL | Soft PPL | Soft KL |
| Unwatermarked | – | – | – | – | 3.86 | 0 | 0 | 0 | 0 | 0 |
| PPL (Soft) | 0 | Bin.$(30, 0.5)$ | 95.9 | 84.5 | 3.85 | $4.3{\cdot}10^{-5}$ | $2.7{\cdot}10^{8}$ | 0.043 | 0.044 | $6.5{\cdot}10^{-7}$ |
| | 0.1 | Bin.$(30, 0.5)$ | 98.8 | 93.9 | 4.12 | $4.9{\cdot}10^{-5}$ | $2.5{\cdot}10^{8}$ | 0.125 | 0.120 | $5.8{\cdot}10^{-7}$ |
| | 0.2 | Bin.$(30, 0.5)$ | 99.1 | 97.4 | 4.60 | $5.4{\cdot}10^{-5}$ | $3.1{\cdot}10^{8}$ | 0.221 | 0.218 | $6.2{\cdot}10^{-7}$ |
| | 0.3 | Bin.$(30, 0.5)$ | 99.5 | 99.0 | 5.10 | $5.9{\cdot}10^{-5}$ | $4.3{\cdot}10^{8}$ | 0.327 | 0.317 | $7.0{\cdot}10^{-7}$ |
| | 0.5 | Bin.$(30, 0.5)$ | 99.8 | 99.6 | 6.15 | $7.0{\cdot}10^{-5}$ | $6.5{\cdot}10^{8}$ | 0.514 | 0.509 | $9.5{\cdot}10^{-7}$ |
| PPL (Hard) | 0 | Bin.$(30, 0.5)$ | 56.8 | 34.2 | 3.75 | $2.4{\cdot}10^{-5}$ | $2.5{\cdot}10^{5}$ | $-0.005$ | $-0.006$ | $3.1{\cdot}10^{-7}$ |
| | 0.1 | Bin.$(30, 0.5)$ | 75.4 | 51.6 | 4.27 | $2.6{\cdot}10^{-5}$ | $1.4{\cdot}10^{6}$ | 0.089 | 0.091 | $2.7{\cdot}10^{-7}$ |
| | 0.2 | Bin.$(30, 0.5)$ | 86.4 | 69.6 | 4.51 | $2.9{\cdot}10^{-5}$ | $4.1{\cdot}10^{6}$ | 0.184 | 0.187 | $3.2{\cdot}10^{-7}$ |
| | 0.5 | Bin.$(30, 0.5)$ | 97.8 | 94.1 | 5.91 | $3.9{\cdot}10^{-5}$ | $1.6{\cdot}10^{7}$ | 0.468 | 0.476 | $6.4{\cdot}10^{-7}$ |
| | 1 | Bin.$(30, 0.5)$ | 99.6 | 99.4 | 10.94 | $5.9{\cdot}10^{-5}$ | $3.1{\cdot}10^{7}$ | 0.925 | 0.939 | $1.8{\cdot}10^{-6}$ |
| | 1.5 | Bin.$(30, 0.5)$ | 99.9 | 99.9 | 22.05 | $8.0{\cdot}10^{-5}$ | $3.9{\cdot}10^{7}$ | 1.346 | 1.358 | $3.7{\cdot}10^{-6}$ |
| AAR/KTH | 0 | Gumbel$(0, 1)$ | 82.3 | 61.4 | 3.73 | $4.2{\cdot}10^{-5}$ | 30.4 | 0.003 | $-0.002$ | $1.1{\cdot}10^{-6}$ |
| | 0.1 | Gumbel$(0, 1)$ | 86.6 | 69.3 | 3.96 | $4.5{\cdot}10^{-5}$ | 108 | 0.057 | 0.053 | $9.2{\cdot}10^{-7}$ |
| | 0.2 | Gumbel$(0, 1)$ | 92.3 | 82.1 | 4.36 | $4.9{\cdot}10^{-5}$ | 504 | 0.135 | 0.119 | $8.5{\cdot}10^{-7}$ |
| | 0.3 | Gumbel$(0, 1)$ | 95.9 | 87.8 | 5.19 | $5.4{\cdot}10^{-5}$ | $1.5{\cdot}10^{3}$ | 0.230 | 0.201 | $9.5{\cdot}10^{-7}$ |
| | 0.4 | Gumbel$(0, 1)$ | 96.8 | 90.8 | 6.07 | $5.9{\cdot}10^{-5}$ | $1.1{\cdot}10^{4}$ | 0.358 | 0.296 | $1.1{\cdot}10^{-6}$ |
| | 0.5 | Gumbel$(0, 1)$ | 98.7 | 95.5 | 16.88 | $6.8{\cdot}10^{-5}$ | $3.7{\cdot}10^{4}$ | 0.585 | 0.400 | $2.2{\cdot}10^{-6}$ |
| KGW | 0.5 | Bern.$(0.5)$ | 8.7 | 1.6 | 3.75 | $1.0{\cdot}10^{-7}$ | 0.0273 | 0.011 | 0.011 | $3.7{\cdot}10^{-8}$ |
| | 1 | Bern.$(0.5)$ | 41.2 | 16.5 | 3.83 | $4.1{\cdot}10^{-7}$ | 0.12 | 0.047 | 0.048 | $5.7{\cdot}10^{-8}$ |
| | 2 | Bern.$(0.5)$ | 92.0 | 77.2 | 4.42 | $1.7{\cdot}10^{-6}$ | 0.699 | 0.188 | 0.191 | $2.5{\cdot}10^{-7}$ |
| | 3 | Bern.$(0.5)$ | 99.1 | 97.3 | 5.39 | $4.0{\cdot}10^{-6}$ | 2.63 | 0.398 | 0.404 | $6.4{\cdot}10^{-7}$ |
| | 4 | Bern.$(0.5)$ | 100.0 | 99.7 | 7.22 | $7.5{\cdot}10^{-6}$ | 8.51 | 0.635 | 0.647 | $1.1{\cdot}10^{-6}$ |
| $\chi^2$ | 0.1 | Bin.$(30, 0.5)$ | 12.0 | 2.6 | 3.66 | $1.5{\cdot}10^{-7}$ | 0.0298 | 0.005 | 0.003 | $3.2{\cdot}10^{-8}$ |
| | 0.5 | Bin.$(30, 0.5)$ | 85.3 | 68.3 | 4.19 | $1.1{\cdot}10^{-5}$ | 0.398 | 0.099 | 0.099 | $1.2{\cdot}10^{-7}$ |
| | 1 | Bin.$(30, 0.5)$ | 97.6 | 92.3 | 4.80 | $2.4{\cdot}10^{-5}$ | 1 | 0.233 | 0.231 | $3.7{\cdot}10^{-7}$ |
| | 1.5 | Bin.$(30, 0.5)$ | 98.8 | 97.5 | 5.59 | $3.3{\cdot}10^{-5}$ | 1.67 | 0.344 | 0.345 | $6.4{\cdot}10^{-7}$ |
| | 2 | Bin.$(30, 0.5)$ | 99.2 | 98.3 | 5.89 | $4.0{\cdot}10^{-5}$ | 2.37 | 0.436 | 0.440 | $8.8{\cdot}10^{-7}$ |
| SynthID | 2 | $m = 15$ | 85.4 | 65.9 | 3.66 | $1.2{\cdot}10^{-5}$ | 1.36 | 0.001 | $-0.002$ | $7.1{\cdot}10^{-8}$ |
| | 3 | $m = 15$ | 97.5 | 93.8 | 4.71 | $3.0{\cdot}10^{-5}$ | 24.1 | 0.212 | 0.214 | $2.9{\cdot}10^{-7}$ |
| | 5 | $m = 10$ | 99.5 | 99.1 | 7.67 | $4.3{\cdot}10^{-5}$ | 152 | 0.611 | 0.605 | $1.0{\cdot}10^{-6}$ |
| | 5 | $m = 15$ | 99.9 | 99.5 | 10.06 | $6.3{\cdot}10^{-5}$ | $2.4{\cdot}10^{3}$ | 0.786 | 0.783 | $1.5{\cdot}10^{-6}$ |
| | 5 | $m = 30$ | 99.7 | 98.4 | 13.34 | $8.7{\cdot}10^{-5}$ | $1.3{\cdot}10^{6}$ | 0.963 | 0.940 | $2.4{\cdot}10^{-6}$ |

Table 7: Full results on MINISTRAL-3-14B and ELI5 (Figure 7): TPR at 1% and 0.1% FPR and mean perplexity.

| | | | Detectability (%) ↑ | | Quality ↓ |
|---|---|---|---|---|---|
| Scheme | $\varepsilon$ | $G$ | TPR@1 | TPR@0.1 | PPL |
| Unwatermarked | – | – | – | – | 4.20 |
| PPL (Soft) | 0 | Bin.$(30, 0.5)$ | 95.8 | 90.2 | 4.39 |
| | 0.1 | Bin.$(30, 0.5)$ | 98.9 | 96.9 | 4.66 |
| | 0.2 | Bin.$(30, 0.5)$ | 99.9 | 99.1 | 5.12 |
| | 0.3 | Bin.$(30, 0.5)$ | 100.0 | 99.5 | 5.58 |
| | 0.5 | Bin.$(30, 0.5)$ | 100.0 | 99.9 | 6.64 |
| PPL (Hard) | 0 | Bin.$(30, 0.5)$ | 66.5 | 44.4 | 4.22 |
| | 0.1 | Bin.$(30, 0.5)$ | 80.5 | 60.4 | 4.64 |
| | 0.2 | Bin.$(30, 0.5)$ | 89.8 | 76.2 | 5.09 |
| | 0.5 | Bin.$(30, 0.5)$ | 98.8 | 96.5 | 6.71 |
| | 1 | Bin.$(30, 0.5)$ | 100.0 | 99.8 | 10.91 |
| | 1.5 | Bin.$(30, 0.5)$ | 100.0 | 100.0 | 18.00 |
| AAR/KTH | 0 | Gumbel$(0, 1)$ | 83.5 | 62.9 | 4.21 |
| | 0.1 | Gumbel$(0, 1)$ | 90.6 | 75.9 | 4.43 |
| | 0.2 | Gumbel$(0, 1)$ | 94.2 | 84.2 | 4.73 |
| | 0.3 | Gumbel$(0, 1)$ | 97.0 | 91.0 | 5.03 |
| | 0.4 | Gumbel$(0, 1)$ | 98.4 | 95.1 | 5.60 |
| | 0.5 | Gumbel$(0, 1)$ | 98.8 | 96.5 | 6.32 |
| KGW | 0.5 | Bern.$(0.5)$ | 10.9 | 1.8 | 4.26 |
| | 1 | Bern.$(0.5)$ | 46.1 | 19.7 | 4.42 |
| | 2 | Bern.$(0.5)$ | 94.7 | 83.4 | 5.01 |
| | 3 | Bern.$(0.5)$ | 99.2 | 98.4 | 6.13 |
| | 4 | Bern.$(0.5)$ | 100.0 | 100.0 | 7.73 |
| $\chi^2$ | 0.1 | Bin.$(30, 0.5)$ | 14.1 | 3.4 | 4.24 |
| | 0.5 | Bin.$(30, 0.5)$ | 91.5 | 74.9 | 4.65 |
| | 1 | Bin.$(30, 0.5)$ | 97.9 | 92.9 | 5.22 |
| | 1.5 | Bin.$(30, 0.5)$ | 99.2 | 97.5 | 5.69 |
| | 2 | Bin.$(30, 0.5)$ | 99.5 | 98.5 | 6.23 |
| SynthID | 1 | Bin.$(30, 0.5)$ | 86.8 | 69.7 | 4.22 |
| | 2 | Bin.$(30, 0.5)$ | 90.2 | 75.2 | 5.38 |

Table 8: Full results on the additional models and datasets (Figure 8 and Table 4): TPR (%) at 1% and 0.1% FPR and mean perplexity. For SynthID, $\varepsilon$ is the number of candidates per tournament and $m$ the number of tournament layers (with Bernoulli$(0.5)$ g-values). (Part 1/2.)

**GEMMA4-12B · ELI5**

| Scheme | $\varepsilon$ | $G$ | TPR@1 | TPR@0.1 | PPL |
|---|---|---|---|---|---|
| PPL (Soft) | 0 | Bin.$(30, 0.5)$ | 60.3 | 33.7 | 6.14 |
| | 0.1 | Bin.$(30, 0.5)$ | 76.8 | 52.9 | 6.24 |
| | 0.2 | Bin.$(30, 0.5)$ | 88.3 | 71.2 | 6.55 |
| | 0.3 | Bin.$(30, 0.5)$ | 95.1 | 86.3 | 6.81 |
| | 0.5 | Bin.$(30, 0.5)$ | 98.5 | 95.7 | 7.44 |
| PPL (Hard) | 0 | Bin.$(30, 0.5)$ | 13.9 | 3.9 | 6.01 |
| | 0.1 | Bin.$(30, 0.5)$ | 24.6 | 9.1 | 6.24 |
| | 0.2 | Bin.$(30, 0.5)$ | 38.6 | 18.0 | 6.56 |
| | 0.5 | Bin.$(30, 0.5)$ | 70.3 | 47.2 | 7.50 |
| | 1 | Bin.$(30, 0.5)$ | 95.2 | 85.2 | 9.41 |
| AAR/KTH | 0 | Gumbel$(0, 1)$ | 30.2 | 10.5 | 5.88 |
| | 0.1 | Gumbel$(0, 1)$ | 39.4 | 14.5 | 5.93 |
| | 0.2 | Gumbel$(0, 1)$ | 44.8 | 20.8 | 6.00 |
| | 0.3 | Gumbel$(0, 1)$ | 53.1 | 26.2 | 6.10 |
| | 0.4 | Gumbel$(0, 1)$ | 61.9 | 33.9 | 6.19 |
| | 0.5 | Gumbel$(0, 1)$ | 68.6 | 41.6 | 6.36 |
| KGW | 0.25 | Gumbel$(0, 1)$ | 3.4 | 0.7 | 5.96 |
| | 0.375 | Gumbel$(0, 1)$ | 7.4 | 1.1 | 6.00 |
| | 0.5 | Gumbel$(0, 1)$ | 14.2 | 3.4 | 6.05 |
| | 0.75 | Gumbel$(0, 1)$ | 36.0 | 11.6 | 6.20 |
| | 1 | Gumbel$(0, 1)$ | 66.2 | 38.7 | 6.60 |
| $\chi^2$ | 0.1 | Bin.$(30, 0.5)$ | 5.5 | 1.2 | 5.92 |
| | 0.5 | Bin.$(30, 0.5)$ | 41.2 | 19.0 | 6.10 |
| | 1 | Bin.$(30, 0.5)$ | 65.0 | 39.6 | 6.24 |
| | 1.5 | Bin.$(30, 0.5)$ | 75.4 | 53.6 | 6.43 |
| SynthID | 2 | $m = 15$ | 34.1 | 13.3 | 6.00 |
| | 3 | $m = 15$ | 62.9 | 38.3 | 6.15 |
| | 5 | $m = 10$ | 86.7 | 70.2 | 6.67 |
| | 5 | $m = 15$ | 87.7 | 72.1 | 6.91 |
| | 5 | $m = 30$ | 81.1 | 60.0 | 7.12 |

**GPT-OSS-20B · ELI5**

| Scheme | $\varepsilon$ | $G$ | TPR@1 | TPR@0.1 | PPL |
|---|---|---|---|---|---|
| Unwatermarked | – | – | – | – | 10.86 |
| PPL (Soft) | 0 | Bin.$(30, 0.5)$ | 99.9 | 99.6 | 10.84 |
| | 0.1 | Bin.$(30, 0.5)$ | 99.9 | 99.9 | 11.62 |
| | 0.2 | Bin.$(30, 0.5)$ | 99.9 | 99.9 | 12.67 |
| | 0.3 | Bin.$(30, 0.5)$ | 100.0 | 100.0 | 13.41 |
| | 0.5 | Bin.$(30, 0.5)$ | 100.0 | 99.9 | 15.28 |
| PPL (Hard) | 0 | Bin.$(30, 0.5)$ | 99.3 | 97.1 | 10.69 |
| | 0.1 | Bin.$(30, 0.5)$ | 99.6 | 99.2 | 11.44 |
| | 0.5 | Bin.$(30, 0.5)$ | 100.0 | 99.9 | 15.42 |
| | 1 | Bin.$(30, 0.5)$ | 100.0 | 100.0 | 21.50 |
| | 1.5 | Bin.$(30, 0.5)$ | 100.0 | 100.0 | 30.02 |
| AAR/KTH | 0 | Gumbel$(0, 1)$ | 99.8 | 99.5 | 10.75 |
| | 0.1 | Gumbel$(0, 1)$ | 99.7 | 99.4 | 11.74 |
| | 0.2 | Gumbel$(0, 1)$ | 99.9 | 99.4 | 13.28 |
| | 0.3 | Gumbel$(0, 1)$ | 100.0 | 99.7 | 14.32 |
| | 0.4 | Gumbel$(0, 1)$ | 99.9 | 99.7 | 16.28 |
| | 0.5 | Gumbel$(0, 1)$ | 100.0 | 100.0 | 18.43 |
| KGW | 0.5 | Bern.$(0.5)$ | 30.4 | 9.8 | 11.10 |
| | 1 | Bern.$(0.5)$ | 88.4 | 67.9 | 11.54 |
| | 2 | Bern.$(0.5)$ | 99.8 | 99.5 | 13.48 |
| | 3 | Bern.$(0.5)$ | 100.0 | 100.0 | 15.77 |
| | 4 | Bern.$(0.5)$ | 100.0 | 100.0 | 18.48 |
| $\chi^2$ | 0.1 | Bin.$(30, 0.5)$ | 43.4 | 17.1 | 10.90 |
| | 0.5 | Bin.$(30, 0.5)$ | 99.8 | 99.6 | 12.16 |
| | 1 | Bin.$(30, 0.5)$ | 100.0 | 99.9 | 14.15 |
| | 1.5 | Bin.$(30, 0.5)$ | 100.0 | 100.0 | 16.04 |
| | 2 | Bin.$(30, 0.5)$ | 100.0 | 100.0 | 17.32 |

Table 9: Full results on the additional models and datasets (Part 2/2, continued from Table 8).

**LLAMA3.1-8B · C4**

| Scheme | $\varepsilon$ | $G$ | TPR@1 | TPR@0.1 | PPL |
|---|---|---|---|---|---|
| PPL (Soft) | 0 | Bin.$(30, 0.5)$ | 99.6 | 99.3 | 7.22 |
| | 0.1 | Bin.$(30, 0.5)$ | 99.7 | 99.5 | 7.84 |
| | 0.2 | Bin.$(30, 0.5)$ | 99.9 | 99.5 | 8.56 |
| | 0.3 | Bin.$(30, 0.5)$ | 99.8 | 99.7 | 9.34 |
| | 0.5 | Bin.$(30, 0.5)$ | 100.0 | 99.8 | 11.35 |
| PPL (Hard) | 0 | Bin.$(30, 0.5)$ | 98.5 | 97.9 | 7.14 |
| | 0.1 | Bin.$(30, 0.5)$ | 99.3 | 98.8 | 7.78 |
| | 0.2 | Bin.$(30, 0.5)$ | 99.7 | 99.5 | 8.50 |
| | 0.5 | Bin.$(30, 0.5)$ | 99.9 | 99.7 | 10.97 |
| | 1 | Bin.$(30, 0.5)$ | 99.9 | 99.8 | 17.42 |
| AAR/KTH | 0 | Gumbel$(0, 1)$ | 99.1 | 98.4 | 7.52 |
| | 0.1 | Gumbel$(0, 1)$ | 99.6 | 98.6 | 9.11 |
| | 0.2 | Gumbel$(0, 1)$ | 99.5 | 99.3 | 11.92 |
| | 0.3 | Gumbel$(0, 1)$ | 99.6 | 99.5 | 19.06 |
| | 0.4 | Gumbel$(0, 1)$ | 99.9 | 99.8 | 130.6 |
| | 0.5 | Gumbel$(0, 1)$ | 100.0 | 99.9 | 772.5 |
| KGW | 0.25 | Gumbel$(0, 1)$ | 57.1 | 26.4 | 7.64 |
| | 0.375 | Gumbel$(0, 1)$ | 89.8 | 70.2 | 8.09 |
| | 0.5 | Gumbel$(0, 1)$ | 98.5 | 95.3 | 8.96 |
| | 0.75 | Gumbel$(0, 1)$ | 99.8 | 99.0 | 11.95 |
| | 1 | Gumbel$(0, 1)$ | 99.9 | 99.8 | 19.10 |
| $\chi^2$ | 0.1 | Bin.$(30, 0.5)$ | 49.6 | 24.2 | 7.40 |
| | 0.5 | Bin.$(30, 0.5)$ | 99.0 | 98.3 | 8.62 |
| | 1 | Bin.$(30, 0.5)$ | 99.6 | 99.6 | 10.18 |
| | 1.5 | Bin.$(30, 0.5)$ | 99.7 | 99.4 | 11.71 |
| SynthID | 2 | $m = 15$ | 99.2 | 98.7 | 7.47 |
| | 3 | $m = 15$ | 99.8 | 99.6 | 10.61 |
| | 5 | $m = 10$ | 99.9 | 99.8 | 18.70 |
| | 5 | $m = 15$ | 99.8 | 99.7 | 23.49 |
| | 5 | $m = 30$ | 99.9 | 99.8 | 26.73 |

**LLAMA3.1-8B · OpenMathInstruct-2**

| Scheme | $\varepsilon$ | $G$ | TPR@1 | TPR@0.1 | PPL |
|---|---|---|---|---|---|
| PPL (Soft) | 0 | Bin.$(30, 0.5)$ | 68.5 | 48.1 | 2.44 |
| | 0.1 | Bin.$(30, 0.5)$ | 78.4 | 64.1 | 2.68 |
| | 0.2 | Bin.$(30, 0.5)$ | 89.0 | 79.3 | 3.05 |
| | 0.3 | Bin.$(30, 0.5)$ | 95.2 | 88.1 | 3.50 |
| | 0.5 | Bin.$(30, 0.5)$ | 99.5 | 98.0 | 4.49 |
| PPL (Hard) | 0 | Bin.$(30, 0.5)$ | 30.7 | 15.0 | 2.25 |
| | 0.1 | Bin.$(30, 0.5)$ | 44.7 | 26.0 | 2.57 |
| | 0.2 | Bin.$(30, 0.5)$ | 57.9 | 40.2 | 2.92 |
| | 0.5 | Bin.$(30, 0.5)$ | 87.4 | 77.7 | 4.21 |
| | 1 | Bin.$(30, 0.5)$ | 98.8 | 97.7 | 8.40 |
| AAR/KTH | 0 | Gumbel$(0, 1)$ | 42.7 | 24.2 | 2.31 |
| | 0.1 | Gumbel$(0, 1)$ | 51.4 | 31.8 | 2.41 |
| | 0.2 | Gumbel$(0, 1)$ | 58.1 | 39.7 | 2.56 |
| | 0.3 | Gumbel$(0, 1)$ | 66.9 | 50.1 | 2.92 |
| | 0.4 | Gumbel$(0, 1)$ | 72.7 | 56.6 | 3.64 |
| | 0.5 | Gumbel$(0, 1)$ | 80.3 | 67.0 | 6.84 |
| KGW | 0.25 | Gumbel$(0, 1)$ | 6.3 | 0.9 | 2.33 |
| | 0.375 | Gumbel$(0, 1)$ | 16.7 | 5.2 | 2.44 |
| | 0.5 | Gumbel$(0, 1)$ | 27.9 | 12.3 | 2.55 |
| | 0.75 | Gumbel$(0, 1)$ | 56.7 | 39.7 | 3.13 |
| | 1 | Gumbel$(0, 1)$ | 79.1 | 68.0 | 5.24 |
| $\chi^2$ | 0.1 | Bin.$(30, 0.5)$ | 7.1 | 1.7 | 2.33 |
| | 0.5 | Bin.$(30, 0.5)$ | 53.1 | 32.5 | 2.54 |
| | 1 | Bin.$(30, 0.5)$ | 73.9 | 57.6 | 2.88 |
| | 1.5 | Bin.$(30, 0.5)$ | 82.8 | 70.6 | 3.40 |
| SynthID | 2 | $m = 15$ | 45.9 | 28.9 | 2.29 |
| | 3 | $m = 15$ | 71.7 | 58.2 | 2.87 |
| | 5 | $m = 10$ | 90.0 | 83.0 | 5.13 |
| | 5 | $m = 15$ | 91.2 | 85.3 | 6.82 |
| | 5 | $m = 30$ | 91.0 | 84.5 | 8.39 |

Table 10: Full results of the score distribution ablation (Figure 6 and Table 3): TPR (%) at 1% FPR, median perplexity, and median of the scheme's own constraint, for each score distribution $G$ and strength $\varepsilon$. The unwatermarked model has a median perplexity of 3.56. (Part 1/2.)

**KGW watermark**

| $G$ | $\varepsilon$ | TPR@1 | PPL | KL |
|---|---|---|---|---|
| Bernoulli | 0.5 | 9.1 | 3.56 | $1.0{\cdot}10^{-7}$ |
| | 1 | 34.9 | 3.68 | $4.1{\cdot}10^{-7}$ |
| | 2 | 91.8 | 4.12 | $1.7{\cdot}10^{-6}$ |
| | 3 | 99.2 | 5.04 | $4.0{\cdot}10^{-6}$ |
| | 4 | 99.8 | 6.55 | $7.5{\cdot}10^{-6}$ |
| Binomial | 0.0167 | 1.5 | 3.55 | $3.5{\cdot}10^{-9}$ |
| | 0.1 | 13.9 | 3.55 | $1.2{\cdot}10^{-7}$ |
| | 0.133 | 26.3 | 3.59 | $2.2{\cdot}10^{-7}$ |
| | 0.25 | 73.5 | 3.84 | $8.0{\cdot}10^{-7}$ |
| | 0.4 | 96.1 | 4.36 | $2.1{\cdot}10^{-6}$ |
| | 0.5 | 99.5 | 4.90 | $3.5{\cdot}10^{-6}$ |
| Gumbel | 0.1 | 2.9 | 3.54 | $2.7{\cdot}10^{-8}$ |
| | 0.25 | 13.6 | 3.60 | $1.8{\cdot}10^{-7}$ |
| | 0.5 | 54.9 | 3.86 | $7.6{\cdot}10^{-7}$ |
| | 0.75 | 89.6 | 4.44 | $2.0{\cdot}10^{-6}$ |
| | 1 | 98.4 | 6.07 | $4.2{\cdot}10^{-6}$ |
| Lognormal | 0.1 | 8.5 | 3.56 | $8.3{\cdot}10^{-8}$ |
| | 0.25 | 61.1 | 4.05 | $8.3{\cdot}10^{-7}$ |
| | 0.3 | 77.0 | 4.41 | $1.4{\cdot}10^{-6}$ |
| | 0.4 | 96.5 | 6.20 | $3.5{\cdot}10^{-6}$ |
| Normal | 0.25 | 12.2 | 3.59 | $1.0{\cdot}10^{-7}$ |
| | 0.5 | 45.0 | 3.71 | $4.1{\cdot}10^{-7}$ |
| | 0.75 | 79.1 | 3.83 | $9.4{\cdot}10^{-7}$ |
| | 1 | 94.0 | 4.19 | $1.7{\cdot}10^{-6}$ |
| | 2 | 99.8 | 9.93 | $9.6{\cdot}10^{-6}$ |
| Uniform | 0.5 | 3.9 | 3.55 | $3.4{\cdot}10^{-8}$ |
| | 1 | 12.1 | 3.58 | $1.3{\cdot}10^{-7}$ |
| | 2 | 48.6 | 3.73 | $5.6{\cdot}10^{-7}$ |
| | 3 | 80.6 | 3.98 | $1.3{\cdot}10^{-6}$ |
| | 4 | 94.8 | 4.38 | $2.3{\cdot}10^{-6}$ |
| | 5 | 99.0 | 4.99 | $3.8{\cdot}10^{-6}$ |
| | 6 | 99.5 | 6.01 | $5.8{\cdot}10^{-6}$ |

**$\chi^2$ watermark**

| $G$ | $\varepsilon$ | TPR@1 | PPL | $\chi^2$ |
|---|---|---|---|---|
| Bernoulli | 1 | 33.0 | 3.56 | 0.0902 |
| | 2 | 72.2 | 3.76 | 0.29 |
| | 2.25 | 79.1 | 4.64 | $1.4{\cdot}10^{4}$ |
| | 2.5 | 87.3 | 5.89 | $1.1{\cdot}10^{5}$ |
| | 3 | 92.9 | 10.19 | $1.7{\cdot}10^{5}$ |
| Binomial | 0.1 | 11.9 | 3.56 | 0.0293 |
| | 0.5 | 84.7 | 3.80 | 0.391 |
| | 1 | 97.3 | 4.29 | 0.97 |
| | 1.5 | 99.1 | 4.63 | 1.6 |
| | 2 | 99.0 | 5.12 | 2.32 |
| Gumbel | 0.1 | 3.3 | 3.51 | $6.8{\cdot}10^{-3}$ |
| | 0.5 | 37.5 | 3.60 | 0.135 |
| | 1 | 72.9 | 3.80 | 0.386 |
| | 1.5 | 88.4 | 4.00 | 0.685 |
| | 2 | 93.6 | 4.23 | 1.01 |
| | 2.5 | 97.0 | 4.50 | 1.37 |
| Lognormal | 0.1 | 5.4 | 3.53 | 0.0168 |
| | 0.5 | 58.2 | 3.67 | 0.276 |
| | 1 | 87.5 | 3.95 | 0.797 |
| | 1.5 | 95.4 | 4.28 | 1.5 |
| | 2 | 98.3 | 4.48 | 2.22 |
| | 2.5 | 98.8 | 4.81 | 3.17 |
| Normal | 0.1 | 3.7 | 3.55 | $4.2{\cdot}10^{-3}$ |
| | 0.5 | 32.9 | 3.58 | 0.0865 |
| | 1 | 73.9 | 3.73 | 0.251 |
| | 1.5 | 86.6 | 3.87 | 0.447 |
| | 2 | 93.7 | 4.09 | 0.644 |
| | 2.5 | 96.2 | 4.23 | 0.862 |
| | 3 | 97.7 | 4.31 | 1.07 |
| Uniform | 1 | 10.8 | 3.51 | 0.0328 |
| | 2 | 38.6 | 3.57 | 0.113 |
| | 3 | 62.1 | 3.68 | 0.209 |
| | 5 | 84.1 | 3.90 | 0.425 |
| | 10 | 96.5 | 4.36 | 1.04 |
| | 15 | 98.9 | 4.83 | 1.68 |

Table 11: Full results of the score distribution ablation (Part 2/2, continued from Table 10).

**PPL (Hard) watermark**

| $G$ | $\varepsilon$ | TPR@1 | PPL | Hard PPL |
|---|---|---|---|---|
| Bernoulli | 0 | 43.1 | 3.51 | −0.001 |
| | 0.1 | 61.9 | 3.85 | 0.095 |
| | 0.2 | 74.0 | 4.09 | 0.191 |
| | 0.5 | 94.5 | 5.43 | 0.481 |
| | 1 | 99.8 | 10.08 | 0.955 |
| | 1.5 | 99.8 | 21.04 | 1.404 |
| Binomial | 0 | 59.8 | 3.52 | −0.004 |
| | 0.1 | 75.4 | 3.81 | 0.091 |
| | 0.2 | 86.4 | 4.09 | 0.186 |
| | 0.5 | 97.8 | 5.34 | 0.470 |
| | 1 | 99.9 | 9.58 | 0.931 |
| | 1.5 | 100.0 | 19.89 | 1.357 |
| Gumbel | 0 | 45.6 | 3.54 | −0.006 |
| | 0.1 | 63.9 | 3.80 | 0.088 |
| | 0.2 | 78.3 | 4.12 | 0.182 |
| | 0.5 | 95.3 | 5.35 | 0.463 |
| | 1 | 99.5 | 9.24 | 0.919 |
| | 1.5 | 99.9 | 19.32 | 1.324 |
| Lognormal | 0 | 49.4 | 3.56 | −0.006 |
| | 0.1 | 65.9 | 3.77 | 0.088 |
| | 0.2 | 80.0 | 4.14 | 0.182 |
| | 0.5 | 96.1 | 5.34 | 0.461 |
| | 1 | 99.7 | 9.73 | 0.915 |
| | 1.5 | 99.8 | 18.64 | 1.327 |
| Normal | 0 | 59.7 | 3.56 | −0.006 |
| | 0.1 | 74.6 | 3.82 | 0.088 |
| | 0.2 | 85.8 | 4.10 | 0.182 |
| | 0.5 | 98.3 | 5.35 | 0.464 |
| | 1 | 99.4 | 9.53 | 0.918 |
| | 1.5 | 99.7 | 18.40 | 1.331 |
| Uniform | 0 | 49.9 | 3.54 | −0.006 |
| | 0.1 | 65.3 | 3.79 | 0.089 |
| | 0.2 | 79.9 | 4.13 | 0.183 |
| | 0.5 | 95.6 | 5.31 | 0.466 |
| | 1 | 99.6 | 9.41 | 0.921 |
| | 1.5 | 99.9 | 18.45 | 1.336 |

**PPL (Soft) watermark**

| $G$ | $\varepsilon$ | TPR@1 | PPL | Soft PPL |
|---|---|---|---|---|
| Bernoulli | 0 | 99.6 | 5.38 | 0.482 |
| | 0.1 | 99.7 | 5.50 | 0.518 |
| | 0.2 | 99.8 | 5.64 | 0.540 |
| | 0.3 | 99.7 | 5.82 | 0.563 |
| | 0.5 | 99.8 | 6.21 | 0.610 |
| Binomial | 0 | 94.8 | 3.65 | 0.044 |
| | 0.1 | 98.8 | 3.88 | 0.121 |
| | 0.2 | 99.4 | 4.19 | 0.219 |
| | 0.3 | 99.5 | 4.59 | 0.317 |
| | 0.5 | 99.6 | 5.66 | 0.509 |
| Gumbel | 0 | 83.9 | 3.66 | 0.050 |
| | 0.1 | 93.6 | 3.95 | 0.128 |
| | 0.2 | 97.9 | 4.24 | 0.223 |
| | 0.3 | 98.6 | 4.73 | 0.324 |
| | 0.5 | 99.5 | 5.66 | 0.518 |
| Lognormal | 0 | 81.2 | 3.83 | 0.073 |
| | 0.1 | 88.7 | 4.09 | 0.145 |
| | 0.2 | 93.7 | 4.46 | 0.238 |
| | 0.3 | 96.6 | 4.94 | 0.337 |
| | 0.5 | 99.1 | 5.92 | 0.525 |
| Normal | 0 | 95.5 | 3.64 | 0.044 |
| | 0.1 | 97.2 | 3.91 | 0.122 |
| | 0.2 | 99.4 | 4.25 | 0.220 |
| | 0.3 | 99.9 | 4.60 | 0.319 |
| | 0.5 | 99.9 | 5.58 | 0.511 |
| Uniform | 0 | 90.9 | 3.69 | 0.041 |
| | 0.1 | 95.8 | 3.87 | 0.118 |
| | 0.2 | 99.1 | 4.23 | 0.219 |
| | 0.3 | 99.1 | 4.62 | 0.314 |
| | 0.5 | 99.8 | 5.61 | 0.508 |

Table 12: Full benchmark results on LLAMA3.1-8B (Figures 9 and 10 and Table 5): TPR at 1% FPR on ELI5 and accuracy on HumanEval (pass@1), GSM8K (exact match), and MBPP (pass@1). For SynthID, $\varepsilon$ is the number of candidates per tournament and $m$ the number of tournament layers (with Bernoulli$(0.5)$ g-values).

| | | | Detect. (%) ↑ | Accuracy (%) ↑ | | | |
|---|---|---|---|---|---|---|---|
| Scheme | $\varepsilon$ | $G$ | TPR@1 | HumanEval | GSM8K | MBPP | Average |
| Unwatermarked | – | – | – | 47.6 | 68.7 | 51.6 | 55.9 |
| PPL (Soft) | 0 | Bin.$(30, 0.5)$ | 95.9 | 47.0 | 66.0 | 49.6 | 54.2 |
| | 0.1 | Bin.$(30, 0.5)$ | 98.8 | 45.1 | 62.2 | 47.4 | 51.6 |
| | 0.2 | Bin.$(30, 0.5)$ | 99.1 | 37.2 | 56.2 | 44.6 | 46.0 |
| | 0.3 | Bin.$(30, 0.5)$ | 99.5 | 34.1 | 52.1 | 38.0 | 41.4 |
| | 0.5 | Bin.$(30, 0.5)$ | 99.8 | 25.0 | 43.1 | 32.8 | 33.6 |
| PPL (Hard) | 0 | Bin.$(30, 0.5)$ | 56.8 | 51.8 | 68.5 | 53.2 | 57.8 |
| | 0.1 | Bin.$(30, 0.5)$ | 75.4 | 45.7 | 62.2 | 47.8 | 51.9 |
| | 0.2 | Bin.$(30, 0.5)$ | 86.4 | 39.0 | 56.5 | 46.2 | 47.2 |
| | 0.5 | Bin.$(30, 0.5)$ | 97.8 | 24.4 | 41.3 | 34.8 | 33.5 |
| | 1 | Bin.$(30, 0.5)$ | 99.6 | 15.2 | 26.3 | 22.6 | 21.4 |
| | 1.5 | Bin.$(30, 0.5)$ | 99.9 | 7.9 | 14.9 | 13.2 | 12.0 |
| AAR/KTH | 0 | Gumbel$(0, 1)$ | 82.3 | 50.6 | 70.7 | 53.2 | 58.2 |
| | 0.1 | Gumbel$(0, 1)$ | 86.6 | 48.2 | 67.9 | 52.2 | 56.1 |
| | 0.2 | Gumbel$(0, 1)$ | 92.3 | 46.3 | 65.0 | 50.6 | 54.0 |
| | 0.3 | Gumbel$(0, 1)$ | 95.9 | 47.6 | 61.9 | 49.2 | 52.9 |
| | 0.4 | Gumbel$(0, 1)$ | 96.8 | 40.9 | 57.1 | 47.8 | 48.6 |
| | 0.5 | Gumbel$(0, 1)$ | 98.7 | 37.8 | 52.0 | 43.8 | 44.5 |
| KGW | 0.5 | Bern.$(0.5)$ | 8.7 | 48.2 | 68.1 | 51.6 | 56.0 |
| | 1 | Bern.$(0.5)$ | 41.2 | 43.3 | 66.9 | 52.8 | 54.3 |
| | 2 | Bern.$(0.5)$ | 92.0 | 41.5 | 62.6 | 48.8 | 51.0 |
| | 3 | Bern.$(0.5)$ | 99.1 | 36.0 | 56.4 | 39.0 | 43.8 |
| | 4 | Bern.$(0.5)$ | 100.0 | 26.2 | 42.2 | 33.6 | 34.0 |
| $\chi^2$ | 0.1 | Bin.$(30, 0.5)$ | 12.0 | 50.6 | 67.5 | 52.8 | 57.0 |
| | 0.5 | Bin.$(30, 0.5)$ | 85.3 | 42.1 | 66.3 | 50.6 | 53.0 |
| | 1 | Bin.$(30, 0.5)$ | 97.6 | 39.6 | 64.4 | 50.2 | 51.4 |
| | 1.5 | Bin.$(30, 0.5)$ | 98.8 | 40.9 | 58.7 | 46.6 | 48.7 |
| | 2 | Bin.$(30, 0.5)$ | 99.2 | 34.1 | 55.8 | 43.0 | 44.3 |
| SynthID | 2 | $m = 30$ | 76.4 | 52.4 | 70.3 | 53.6 | 58.8 |
| | 3 | $m = 30$ | 97.5 | 43.3 | 62.2 | 47.6 | 51.0 |
| | 4 | $m = 30$ | 97.5 | 34.1 | 47.2 | 39.4 | 40.3 |
| | 5 | $m = 30$ | 99.7 | 25.0 | 34.7 | 28.2 | 29.3 |

Table 13: Full robustness results (Figures 11 and 12): TPR at 1% FPR after word-level edits of increasing strength and after DIPPER paraphrasing (order diversity 0). All schemes are at matched quality (PPL $\approx 3.5$). We **bold** the best scheme in each row.

| | TPR@1 (%) ↑ | | | | | |
|---|---|---|---|---|---|---|
| Attack strength | PPL (Soft) | PPL (Hard) | AAR/KTH | KGW | $\chi^2$ | SynthID |
| Unattacked | **99.1** | 72.0 | 91.9 | 13.7 | 19.0 | 93.5 |
| **Word deletion** | | | | | | |
| 10% of words | **75.7** | 29.7 | 55.4 | 3.2 | 6.1 | 50.9 |
| 20% of words | **42.0** | 13.9 | 27.1 | 2.6 | 3.1 | 21.8 |
| 30% of words | **18.0** | 5.7 | 9.9 | 2.0 | 2.4 | 10.8 |
| 40% of words | **7.8** | 2.9 | 3.5 | 1.2 | 0.7 | 4.9 |
| 50% of words | **3.5** | 2.5 | 2.7 | 1.3 | 1.2 | 2.0 |
| **Synonym substitution** | | | | | | |
| 10% of words | **77.0** | 28.8 | 57.9 | 4.0 | 5.2 | 52.6 |
| 20% of words | **43.9** | 15.0 | 31.6 | 2.3 | 3.9 | 24.9 |
| 30% of words | **19.4** | 5.8 | 15.2 | 1.9 | 2.1 | 10.8 |
| 40% of words | **8.7** | 2.3 | 7.2 | 1.4 | 1.7 | 6.3 |
| 50% of words | 4.8 | 1.8 | **5.0** | 0.9 | 1.1 | 3.0 |
| **Context-aware substitution** | | | | | | |
| 10% of words | **88.8** | 43.0 | 75.6 | 4.5 | 7.3 | 71.2 |
| 20% of words | **78.0** | 30.7 | 61.3 | 3.4 | 5.3 | 56.7 |
| 30% of words | **67.0** | 22.6 | 46.5 | 2.6 | 4.7 | 41.2 |
| 40% of words | **49.0** | 16.1 | 33.1 | 2.8 | 5.0 | 29.0 |
| 50% of words | **35.3** | 10.6 | 22.5 | 1.6 | 3.2 | 20.1 |
| **DIPPER paraphrasing** | | | | | | |
| Lexical diversity 20 | **47.8** | 18.9 | 30.6 | 3.2 | 5.6 | 32.0 |
| Lexical diversity 40 | **15.3** | 6.1 | 8.9 | 2.2 | 3.0 | 9.9 |
| Lexical diversity 60 | **3.5** | 2.3 | 2.7 | 1.5 | 2.1 | 2.9 |
| Lexical diversity 80 | 0.8 | **1.9** | 0.9 | 1.1 | 0.9 | **1.9** |
| Lexical diversity 100 | 1.0 | 0.8 | **1.4** | 0.5 | 0.5 | **1.4** |

Table 14: Full results of Figure 13: TPR at several FPRs for the configuration of each scheme whose mean perplexity is closest to 3.5 (low distortion) and 4.5 (high distortion), on 1,000 replies. We **bold** the best scheme in each column and regime. For SynthID, $\varepsilon$ is the number of candidates per tournament and $m$ the number of tournament layers (with Bernoulli(0.5) g-values).

| | | | | TPR (%) at FPR ↑ | | | |
|---|---|---|---|---|---|---|---|
| Scheme | $\varepsilon$ | $G$ | PPL | 10% | 5% | 1% | 0.1% |
| **Low distortion (PPL $\approx 3.5$)** | | | | | | | |
| PPL (Soft) | 0 | Bin.$(30, 0.5)$ | 3.85 | **99.1** | **98.4** | **95.9** | **84.5** |
| PPL (Hard) | 0 | Bin.$(30, 0.5)$ | 3.75 | 84.5 | 76.3 | 56.8 | 34.2 |
| AAR/KTH | 0 | Gumbel$(0, 1)$ | 3.73 | 95.7 | 93.0 | 82.3 | 61.4 |
| KGW | 0.5 | Bern.$(0.5)$ | 3.75 | 39.0 | 24.6 | 8.7 | 1.6 |
| $\chi^2$ | 0.1 | Bin.$(30, 0.5)$ | 3.66 | 44.9 | 32.6 | 12.0 | 2.6 |
| SynthID | 2 | $m = 15$ | 3.66 | 97.1 | 95.6 | 85.4 | 65.9 |
| **High distortion (PPL $\approx 4.5$)** | | | | | | | |
| PPL (Soft) | 0.2 | Bin.$(30, 0.5)$ | 4.60 | **99.5** | **99.2** | **99.1** | **97.4** |
| PPL (Hard) | 0.2 | Bin.$(30, 0.5)$ | 4.51 | 97.0 | 94.0 | 86.4 | 69.6 |
| AAR/KTH | 0.2 | Gumbel$(0, 1)$ | 4.36 | 98.8 | 97.7 | 92.3 | 82.1 |
| KGW | 2 | Bern.$(0.5)$ | 4.42 | 98.6 | 97.5 | 92.0 | 77.2 |
| $\chi^2$ | 1 | Bin.$(30, 0.5)$ | 4.80 | 99.3 | 99.1 | 97.6 | 92.3 |
| SynthID | 4 | $m = 5$ | 4.61 | 99.4 | 98.8 | 97.5 | 93.9 |

Table 15: Full results of Figure 4: for each scheme, the smallest $\varepsilon$ reaching a TPR@1 of at least 95%, its trigram Self-BLEU and its $\log_{10}$ perplexity (mean $\pm$ standard deviation over prompts).

| Scheme | $\varepsilon$ | Self-BLEU ↓ | $\log_{10}$ PPL ↓ |
|---|---|---|---|
| PPL (Soft) | 0 | $0.934 \pm 0.079$ | $0.569 \pm 0.109$ |
| PPL (Hard) | 0.5 | $0.786 \pm 0.091$ | $0.742 \pm 0.149$ |
| AAR/KTH | 0.3 | $0.974 \pm 0.092$ | $0.654 \pm 0.171$ |
| KGW | 3 | $0.878 \pm 0.066$ | $0.717 \pm 0.153$ |
| $\chi^2$ | 1 | $0.825 \pm 0.086$ | $0.641 \pm 0.134$ |
| SynthID | 5 | $0.874 \pm 0.100$ | $0.638 \pm 0.140$ |

# I LICENSES FOR EXISTING ASSETS

We used the following existing assets in the experiments and evaluations.

**Datasets**

- ELI5 (KILT/Hugging Face distribution): MIT License.
- HumanEval: MIT License.
- MBPP: CC BY 4.0 International License.
- GSM8K: MIT License.
- C4 (`realnewslike`): ODC-BY License.
- OpenMathInstruct-2: CC BY 4.0 International License.

**Models**

- Llama3.1-8B-Instruct: Llama 3.1 Community License.
- Qwen3-30B-A3B: Apache License 2.0.
- Ministral-3-14B-Instruct: Apache License 2.0.
- Gemma4-12B-it: Apache License 2.0.
- GPT-OSS-20B: Apache License 2.0.
- BERT base uncased: Apache License 2.0.
- Dipper paraphraser XXL: Apache License 2.0.