Optimizing Watermarks for Large Language Models

Bram Wouters [1]

Abstract

With the rise of large language models (LLMs) and concerns about potential misuse, watermarks for generative LLMs have recently attracted much attention. An important aspect of such watermarks is the trade-off between their identifiability and their impact on the quality of the generated text. This paper introduces a systematic approach to this trade-off in terms of a multi-objective optimization problem. For a large class of robust, efficient watermarks, the associated Pareto optimal solutions are identified and shown to outperform existing robust, efficient watermarks.

1. Introduction

In recent years, transformer-based LLMs have proven to be remarkably powerful. Their societal impact is potentially enormous. As a consequence of their rapid rise, concerns about potential misuse have been raised. One can think of, for example, plagiarism (Meyer et al., 2023), online propaganda (Goldstein et al., 2023), examination in education (Milano et al., 2023), misinformation (Vincent, 2022) and copyright infringement (Rillig et al., 2023). One possible strategy to partially address these concerns is to ensure that LLM-generated text can be algorithmically distinguished from human-generated text by means of a watermark.

Initialized by Aaronson (2022) and the seminal work of Kirchenbauer et al. (2023), the idea of watermarking LLM-generated text has attracted much attention, in both the scientific field (see Section 5 for an overview) and the industry (Bartz & Hu, 2023). Generally speaking, the process of text generation by an LLM would be adjusted in a controllable manner. Based on the generated text, a detector with knowledge of the watermarking strategy is then able to identify a text as generated by an LLM. This is usually done in the form of a hypothesis test, where the null hypothesis is that the text has been generated by a human being.

This paper largely follows the watermarking strategy of Kirchenbauer et al. (2023). Causal LLMs typically generate text word-by-word. Before a word is generated, the complete vocabulary of the LLM is split in two disjunct lists labelled *green* and *red*. This split is pseudo-random, where the seed is determined by the previous word(s). Green-list words are then sampled with a higher probability than the original LLM prescribes, and red-list words with a lower probability. A detector with knowledge of the pseudo-random green-red split can count the number of green-list words in a text. If this number is larger than one would expect from a text generated without knowledge of the green-red split (e.g., a human-generated text), the null hypothesis is rejected and the text is attributed to an LLM.

There are many perspectives of what a good watermark for an LLM comprises of, and even their usefulness altogether is under debate (Sadasivan et al., 2023; Jiang et al., 2023; Zhang et al., 2023). Roughly speaking, the quality of a watermark is assessed along four axes. A watermark must be

- **identifiable**, meaning that a detector is able to correctly identify the generator (LLM vs. human) of a text.
- **stealthy**, meaning that a watermark does not noticeably change the quality of the generated text.
- **robust** against (moderate) post-generation adjustments of the text that could obfuscate the watermark.
- **efficient** at generation and detection time, i.e., without the need for computationally costly processes.

This paper focusses on the trade-off between identifiability and stealthiness of watermarks. The probability that the hypothesis test of the detector draws the correct conclusion increases when the watermark more strongly promotes green-list tokens. However, enforcing green-list tokens too strongly can degrade the quality of the text in unacceptable manners. We will refer to this trade-off between the quality of test and text as the *test-text trade-off*.

Our contribution. For a large class of robust, efficient watermarks based on the green-red split of the vocabulary, we translate the test-text trade-off into a multi-objective optimization problem and identify the associated Pareto optimal solutions. We empirically validate the optimality of

[1] University of Amsterdam. Correspondence to: Bram Wouters <b.m.wouters@uva.nl>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

the solutions and show that they outperform existing proposals of robust, efficient watermarks (Kirchenbauer et al., 2023; Kuditipudi et al., 2024; Wu et al., 2023) with respect to the test-text trade-off.[*] The contribution of this paper is therefore twofold. To the best of our knowledge, this is the first systematic approach to optimizing the trade-off between identifiability and stealthiness of watermarks for LLMs. Secondly, since we optimize over a large class of robust, efficient watermarks, we believe that the optimal watermarks introduced in this paper have an excellent standing with respect to the four criteria for good watermarks for LLMs.

2. Watermarks for LLMs

In this section a class of watermarks is defined, over which the test-text trade-off will be optimized. The class can be seen as a generalization of the watermark introduced by Kirchenbauer et al. (2023), based on a green-red split of the vocabulary.

In the context of LLMs, a text is typically represented as a sequence of tokens. Consider the sequence of random variables $V_1, V_2, \ldots, V_T$, where V_t corresponds to the token at position t. They have support set $\mathcal{V}$ of size $N = |\mathcal{V}|$, which is the vocabulary of the LLM. In addition, there is the prompt $V_{:1} = (V_0, V_{-1}, V_{-2}, \ldots)$, whose length is left unspecified because most modern causal LLMs do not require a fixed prompt length. The joint probability mass function (pmf) is

$$\mathrm{P}[V_T, V_{T-1}, \ldots V_1, V_{:1}] = \prod_{t=1}^{T} \mathrm{P}[V_t|V_{:t}] \times \mathrm{P}[V_{:1}]\,, \tag{1}$$

where the prefix $V_{:t}$ is the subsequence of tokens prior to position t, including the prompt. The causal LLM specifies $\mathrm{P}[V_t|V_{:t}]$, whereas $\mathrm{P}[V_{:1}]$ represents the distribution of the text prompts under consideration. Text generation occurs token-by-token through sampling from the conditional pmf $\mathrm{P}[V_t|V_{:t}]$.

Before defining the watermark, we introduce a function $g_\gamma : \mathcal{V}^* \to \Theta$, where $\mathcal{V}^*$ is the space of all possible prefixes of arbitrary length and Θ is the space of all subsets of size $\lfloor \gamma N \rfloor$ of the vocabulary $\mathcal{V}$. The hyperparameter $\gamma \in (0, 1)$ is fixed. Given a prefix $v_{:t} \in \mathcal{V}^*$, the set $\mathcal{G}_t = g_\gamma(v_{:t})$ contains the so-called green-list tokens. The tokens in the complement $\mathcal{V} \setminus \mathcal{G}_t$ are the red-list tokens. This partitioning of the vocabulary is pseudo-random, with a seed determined by a hash of $v_{:t}$ and a key. The detector of the watermark has the key and is therefore able to reconstruct the list of green tokens for each position t in the sequence.

This paper considers a large class of watermarks, defined by the conditional probability distribution

$$\tilde{\mathrm{P}}[V_t|V_{:t}] = \mathrm{P}[V_t|V_{:t}] \times \begin{cases} 1 + \frac{\Delta(p_t,\mathcal{G}_t)}{\Gamma_t} & \text{if } V_t \in \mathcal{G}_t, \\ 1 - \frac{\Delta(p_t,\mathcal{G}_t)}{1-\Gamma_t} & \text{if } V_t \notin \mathcal{G}_t, \end{cases} \tag{2}$$

where p_t is the function $p_t(v) = \mathrm{P}[V_t = v|V_{:t}]$ for $v \in \mathcal{V}$, representing the conditional pmf of the LLM at position t, and $\Gamma_t = \mathrm{P}[V_t \in \mathcal{G}_t|V_{:t}]$ is the conditional probability that token t is a green-list token. A watermark is specified by a so-called *shift function* $\Delta : \Xi \times \Theta \to [0, 1]$, where Ξ is the space of all pmfs over the vocabulary $\mathcal{V}$. By demanding that $\Delta(p_t, \mathcal{G}_t) \geq 0$, the shift function increases green-list probabilities and decreases red-list probabilities. To be concrete, $\tilde{\mathrm{P}}[V_t \in \mathcal{G}_t|V_{:t}] = \Gamma_t + \Delta(p_t, \mathcal{G}_t)$, i.e., the shift function is the increase due to the watermark of the conditional probability that a token is on the green list. For $\tilde{\mathrm{P}}[V_t|V_{:t}]$ to be a valid conditional pmf, we also must demand that $\Delta(p_t, \mathcal{G}_t) \leq 1 - \Gamma_t$, where it is important to realize that Γ_t, p_t and $\mathcal{G}_t$ are all functions of the prefix $V_{:t}$. A watermarked LLM generates text by sampling from $\tilde{\mathrm{P}}[V_t|V_{:t}]$.

We believe that the watermarks under consideration here, defined by Equation (2), have two important conceptual benefits in terms of their simplicity. First of all, the shift function $\Delta(\cdot, \cdot)$ is not a function of V_t and therefore the watermark does not alter the relative probabilities among green-list tokens, i.e., $\tilde{\mathrm{P}}[V_t|V_{:t}, V_t \in \mathcal{G}_t] = \mathrm{P}[V_t|V_{:t}, V_t \in \mathcal{G}_t]$. In other words, the watermark rescales the probabilities of all green-list tokens by the same factor, and vice versa for red-list tokens. In this sense, the change to the conditional probability distribution due to the watermarks is minimal. Secondly, the alteration of the conditional probabilities of the LLM due to the watermark is solely determined by the conditional probabilities themselves. This must be contrasted by several recent proposals for watermarks that also use a green-red split of the vocabulary, whose watermarking strategy aims to address the test-text trade-off by means of an external model (Fang et al., 2023; Li et al., 2023b; Chen et al., 2024) and/or a word similarity score defined by a metric on the embedding space of word vectors (Fu et al., 2024; Chen et al., 2024).

2.1. The KGW Watermark

Arguably the simplest example of the class of watermarks defined by Equation (2) is the so-called *hard* watermark, $\Delta_{\text{HARD}}(p_t, \mathcal{G}_t) = 1 - \Gamma_t$, implying that $\tilde{\mathrm{P}}[V_t \in \mathcal{G}_t|V_{:t}] = 1$. Green-list tokens are generated with probability one. This watermark is maximally strong, but at the same time impacts the text quality in an unacceptable manner (Kirchenbauer et al., 2023).

This is mitigated by the introduction of what we call the KGW watermark, after the first three author names of Kirchenbauer et al. (2023). Green-list logits are shifted

[*] Code is available at https://github.com/brwo/optimizing-watermarks.

by a watermark parameters $\delta \geq 0$, while red-list logits are left unaltered,

$$\tilde{\mathrm{P}}_{\mathrm{KGW}}[V_t|V_{:t}] = \frac{\exp[\ell(V_t|V_{:t}) + \delta\, \mathrm{I}_{\mathcal{G}_t}(V_t)]}{\sum_{V_t' \in \mathcal{V}} \exp[\ell(V_t'|V_{:t}) + \delta\, \mathrm{I}_{\mathcal{G}_t}(V_t')]},$$

where $\ell(V_t|V_{:t})$ are the logits of the original LLM and $\mathrm{I}_{\mathcal{G}_t}(\cdot)$ is an indicator function. As already mentioned, the KGW watermark is a member of the class of watermarks defined by Equation (2), with

$$\Delta_{\mathrm{KGW}}(p_t, \mathcal{G}_t) = \frac{\Gamma_t(1-\Gamma_t)(e^\delta - 1)}{1 - \Gamma_t + \Gamma_t e^\delta}.$$

The parameter δ controls the test-text trade-off, where a large δ corresponds to a high watermark identifiability.

The logic behind the KGW watermark is that green-list tokens are only substantially favored if this does not hurt text quality. Selecting a green-list token would hurt text quality if all meaningfully probable tokens are on the red list. But these tokens have much higher logits and a moderate shift δ of green-list logits will not change that. We emphasize that this particular choice of watermark is based on a heuristic argument regarding the test-text trade-off, rather than an optimization objective.

The KGW watermark is generally considered to be robust (Shi et al., 2024; Kirchenbauer et al., 2024; Piet et al., 2023) and efficient (Wu et al., 2023). Other watermarks defined by Equation (2) only differ from the KGW watermark through the shift function, which does not impact robustness. Instead, robustness of watermarks based on a green-red split is typically determined by the choice of the green-list generator g_γ (Liu et al., 2024b). And unless the shift function is computationally expensive, which will not be the case in the applications discussed in this paper, all watermarks defined by Equation (2) have a comparable efficiency. We therefore conclude that the watermarks introduced in this paper can be considered robust and efficient.

3. Optimizing Watermarks

Optimization of the test-text trade-off requires a precise definition of both test and text quality. A simple criterion for a good test is a high number of generated green-list tokens, compared to the baseline of the non-watermarked LLM. Let N_g be the number of green-list tokens in the sequence $V_1, V_2, \ldots, V_T$. The expected number of green-list tokens shifts, as a consequence of the watermark, by $\Delta N_g = \tilde{\mathrm{E}}[N_g] - \mathrm{E}[N_g]$, where $\mathrm{E}[\cdot]$ is the expectation with respect to the joint pmf of Equation (1) and $\tilde{\mathrm{E}}[\cdot]$ is the watermarked counterpart. It follows that

$$\Delta N_g = \sum_{t=1}^{T} \tilde{\mathrm{E}}[\Delta(p_t, \mathcal{G}_t)]\,, \qquad (3)$$

as the shift function is the increase (due to the watermark) of the probability that the token is a green-list token.

One common measure for text quality of an LLM is the perplexity, which is the exponential of the negative (normalized) log-likelihood. We consider the log-perplexity

$$\log \mathrm{PPL} = -\frac{1}{T}\sum_{t=1}^{T} \log \mathrm{P}[V_t|V_{:t}],$$

and note that a high text quality corresponds to a low log-perplexity. The shift in expected log-perplexity due to the watermark, $\Delta \log \mathrm{PPL} = \tilde{\mathrm{E}}[\log \mathrm{PPL}] - \mathrm{E}[\log \mathrm{PPL}]$, is given by

$$\Delta \log \mathrm{PPL} = \frac{1}{T}\sum_{t=1}^{T} \tilde{\mathrm{E}}[\Delta(p_t, \mathcal{G}_t) B(p_t, \mathcal{G}_t)]\,, \qquad (4)$$

where

$$B(p_t, \mathcal{G}_t) = \sum_{v \in \mathcal{V}} \frac{\Gamma_t - \mathrm{I}_{\mathcal{G}_t}(v)}{\Gamma_t(1-\Gamma_t)} p_t(v) \log p_t(v).$$

This quantity $B(p_t, \mathcal{G}_t)$ is the expected rate of change of the log-perplexity, given the prefix $V_{:t}$, due to a shift in the conditional probability that the token is a green-list token. Roughly speaking, $B(p_t, \mathcal{G}_t)$ is large when there are no or few green-list tokens with a (relatively) large probability. It should be interpreted as the expected damage that promoting green-list tokens has on the text quality. Equations (3) and (4) are derived under the mild assumption that expectations are unaffected by watermark-induced changes in the distribution of the prefix $V_{:t}$. For details, see Appendix A.

We are now in a position to find a watermark that optimizes the text-test trade-off. Let Υ be the set of all shift functions $\Delta(\cdot,\cdot)$, as defined in Section 2, representing the class of watermarks defined in Equation 2. The aim to maximize test quality and simultaneously minimize a decrease in text quality translates into the multi-objective optimization problem

$$\max_{\Delta \in \Upsilon} \Delta N_g \qquad \text{and} \qquad \min_{\Delta \in \Upsilon} \Delta \log \mathrm{PPL}, \qquad (5)$$

which has Pareto optimal solutions parametrized by $\beta \geq 0$,

$$\Delta_{\mathrm{OPT}}(p_t, \mathcal{G}_t) = \begin{cases} 1 - \Gamma_t & \text{if} \quad B(p_t, \mathcal{G}_t) \leq \beta, \\ 0 & \text{if} \quad B(p_t, \mathcal{G}_t) > \beta. \end{cases} \qquad (6)$$

We will call this the OPT watermark, or simply OPT. For a token at position t there are two options. If the expected damage to the text quality is small, at most β, then the watermark is maximally enforced by generating a green-list token with probability one. Otherwise, no watermark is imposed and the token is sampled from the original LLM. In other words, tokens that are expected to damage the text quality the least are maximally watermarked before other tokens get any watermark at all.

For an intuitive explanation and a formal proof of the Pareto optimality of the OPT watermark, see Appendix A.3. We emphasize that no additional assumptions about the joint pmf in Equation (1) are needed, as the sums of Equations (3) and (4) can be maximized/minimized token-by-token.

Note that maximally watermarking tokens for which $B(p_t, \mathcal{G}_t) < 0$ is actually favorable for the text quality, because you make the sampling more greedy towards green-list tokens and this (potentially) decreases the perplexity. In fact, in the context of LLMs without watermarks greedy sampling minimizes $\mathrm{E}[\log \mathrm{PPL}]$.

3.1. Other Optimized Watermarks

The choice of optimization objectives in Equation (5) is not unique. Different objectives could lead to different optimal watermarks. When a detector performs a hypothesis test based on a sequence of size T, it will typically count the number of green-list words N_g and reject the null hypothesis of a human-generated text if $N_g \geq n^*$. Here, n^* is set beforehand and corresponds to a false-positive rate $\alpha^* = \mathrm{P}[N_g \geq n^*]$. This is the probability of falsely attributing a text to an LLM (type-I error). The quality of the test is then commonly quantified as the power/sensitivity $\pi_{n^*} = \tilde{\mathrm{P}}[N_g \geq n^*]$, i.e., the probability of correctly identifying an LLM-generated text.

Suppose we want to maximize the power of the test for the class of watermarks of Equation (2), i.e., consider the multi-objective optimization problem

$$\max_{\Delta \in \Upsilon} \pi_{n^*} \qquad \text{and} \qquad \min_{\Delta \in \Upsilon} \Delta \log \mathrm{PPL}. \tag{7}$$

The OPT watermark defined in Equation (6) is also Pareto optimal for this optimization problem, provided the following assumptions hold for all t and all $t' \neq t$:

(i) the green-red split is unbiased, $\mathrm{E}[\Gamma_t] = \gamma$,

(ii) the events $V_t \in \mathcal{G}_t$ and $V_{t'} \in \mathcal{G}_{t'}$, which can be seen as Bernoulli random variables, are independent and identically distributed for watermarked text.

The number of green-list tokens is then binomially distributed, $N_g \sim \mathrm{BIN}(T, \gamma + \tilde{\mathrm{E}}[\Delta(p_t, \mathcal{G}_t)])$, and this means that maximizing the power of the test is equivalent to maximizing ΔN_g (see Appendix A.1 for details).

It should be stressed that the required assumptions are implicitly made throughout the literature about watermarks for LLMs, whenever test quality is measured in terms of a z-score, p-value or power of a test, as this requires that N_g is binomially distributed. We also emphasize that the assumptions are only about the events of a token being a green-list token, and not about the distributions of the tokens themselves. In particular, we do not assume that the tokens $V_1, \ldots, V_T$ are independent and/or identically distributed. The validity of the assumptions is further investigated in Section 4.3.

To show that our approach can lead to different optimal watermarks, consider the following possible alternative objective for text quality: minimize the expectation of

$$\frac{1}{T}\sum_{t=1}^{T}(-\log \mathrm{P}[V_t|V_{:t}])^2 = \log \mathrm{PPL}^2 + \frac{1}{T}\sum_{t=1}^{T}(-\log \mathrm{P}[V_t|V_{:t}] - \log \mathrm{PPL})^2,$$

which can be interpreted as the bias with respect to zero (squared) plus the variance. The idea behind this objective is that it seeks to reduce the overall perplexity of a sequence, but also large deviations from this overall perplexity at the level of individual tokens. When simultaneously maximizing the power of the test, the Pareto optimal watermark is now

$$\Delta_{\mathrm{OPT}'}(p_t, \mathcal{G}_t) = \begin{cases} 1 - \Gamma_t & \text{if} \quad B'(p_t, \mathcal{G}_t) \leq \beta', \\ 0 & \text{if} \quad B'(p_t, \mathcal{G}_t) > \beta', \end{cases} \tag{8}$$

parametrized by $\beta' \geq 0$, where

$$B'(p_t, \mathcal{G}_t) = \sum_{v \in \mathcal{V}} \frac{\mathrm{I}_{\mathcal{G}_t}(v) - \Gamma_t}{\Gamma_t(1 - \Gamma_t)} p_t(v)[\log p_t(v)]^2.$$

We will refer to this watermark as OPT′.

4. Experiments

The test-text trade-off is the main focus of this paper. We start by analyzing how the OPT watermark performs in this respect against a baseline of existing proposals for efficient, robust watermarks.

In the subsequent three sections we discuss potential limitations of the general idea of using Pareto optimal watermarks, and the OPT watermark in particular. They can be read as cautionary tales. In Section 4.2 we give two examples of the fact that optimality with respect to one metric for text quality does not necessarily generalize to other metrics. We then investigate in Section 4.3 the validity of the assumptions that were needed to derive the OPT watermark. Finally, in Section 4.4 we present a novel analysis of the effect of watermarking on text diversity.

4.1. Comparing the Test-Text Trade-off of Watermarks

Our experimental setup largely follows Kirchenbauer et al. (2023). From the C4 dataset (Raffel et al., 2020) a sample of 500 (news) articles is drawn randomly. For each text the first (at most) 200 tokens serve as prompt, while the rest is discarded. Based on a prompt, 64 sequences of length $T = 30$ are generated by means of the OPT-1.3B causal LLM (Zhang et al., 2022), or by a watermarked version thereof. With this setup both the prefix $V_{:1}$ and the watermarked sequence $V_1, \dots, V_T$ are sampled. The fact that we restrict ourselves to $T = 30$ is not a limitation. Under the assumption that the tokens $V_1, \dots, V_T$ are identically distributed, the results obtained for sequence length $T = 30$ are easily extendable to larger sequences. Following Kirchenbauer et al. (2023), we let the list of green tokens for position t be determined by only the token at position $t-1$, i.e., $\mathcal{G}_t = g_\gamma(v_{t-1})$. This may not be optimal in the trade-off between tampering-resistancy and invisibility (Liu et al., 2024b), but we consider this outside the scope of this paper. For more details about the setup, see Appendix B.

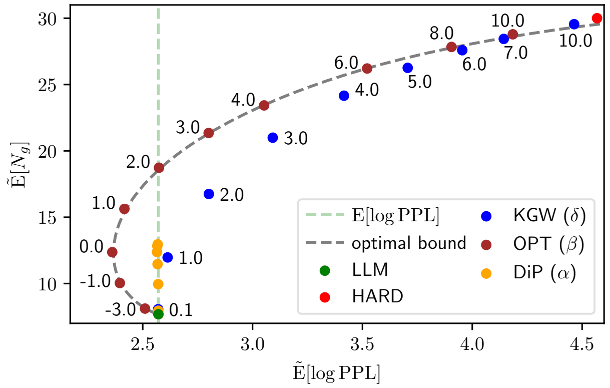

Figure 1. Test quality, measured as the expected number of green-list tokens, versus text quality, measured as the expected log-perplexity, for different watermarks. For DiP we have taken parameter values $\alpha = 0.01, 0.1, 0.2, 0.3, 0.4, 0.5$. As reference, the original language model without watermark is also included (LLM). Also shown is the Pareto optimal bound. Error bars (vertical and horizontal) are omitted, as they are never larger than the marker sizes.

Figures 1 and 2 contain the main empirical results of this paper. Figure 1 shows test quality, measured in terms of the expected number of green-list tokens, versus text quality, measured in terms of the expected log-perplexity. In Figure 2 test quality is measured in terms of the power of the test. In addition to the hard and KGW watermarks, two more robust and (relatively) efficient watermarks are included in the baseline: the distribution-preserving watermark from Wu et al. (2023), with hyperparameter α, and the inverse-transform-sampling watermark from Kuditipudi et al. (2024). We will refer to them as DiP and ITS, respectively. The latter is only shown in Figure 2, as it is not based on a green-red split of the vocabulary and therefore does not have an expected number of green-list tokens. Both these watermarks have the merit of being distortion free, meaning that the watermark does not alter the sampling distribution of the LLM, when averaged over the randomness of the watermark key.

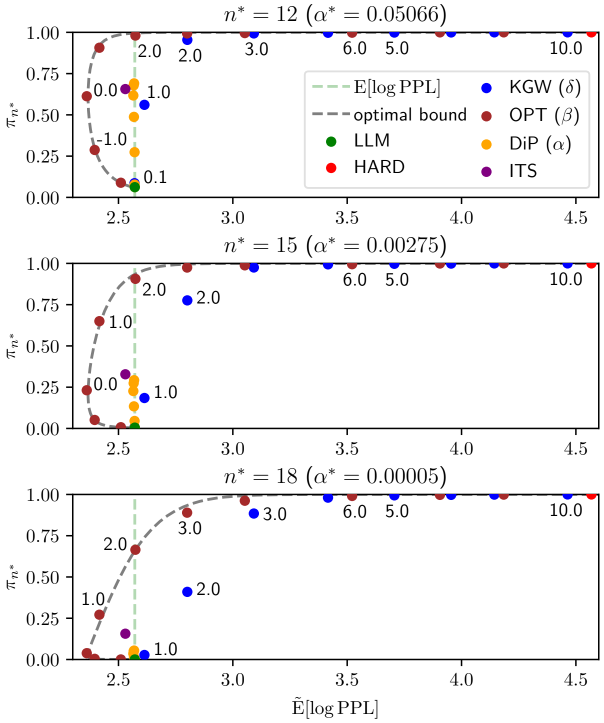

Figure 2. Similar to Figure 1, but now with test quality measured as the power of the test. The panels correspond to different tests ($n^* = 12, 15, 18$) and false-positive rates (α^*). Also included is the ITS watermark.

For the purpose of a fair comparison, we deliberately do not include the watermark from Kuditipudi et al. (2024) that is based on exponential minimum sampling. It is known to incur significant computational costs at detection time and therefore cannot be considered as efficient (Wu et al., 2023; Gu et al., 2024). We also exclude from our analysis the two watermarks of Hu et al. (2024), referred to as δ- and γ−reweighting. In the generation phase they are equivalent to ITS and DiP at parameter value $\alpha = 0.5$, respectively. However, their detectors make use of the weights of the LLM, which might not be available in practice. It also makes them (relatively) inefficient (Kuditipudi et al., 2024).

Figures 1 and 2 show that OPT outperforms KGW in terms of the test-text trade-off, which was expected because of the Pareto optimality of OPT. It also outperforms DiP and ITS. This was not obvious beforehand, as the two watermarks are not in the class of watermarks defined by Equation (2).

Figure 1 also exhibits the curve of $\mathrm{E}[\Delta_{\mathrm{OPT}}(p_t, \mathcal{G}_t)T]$ against $\mathrm{E}[\Delta_{\mathrm{OPT}}(p_t, \mathcal{G}_t)B(p_t, \mathcal{G}_t)]$, parametrized by β. Under the assumptions that $V_1, V_2, \ldots, V_T$ are identically distributed and that the distribution of $(p_t, \mathcal{G}_t)$ does not shift due to the presence of a watermark, this represents the Pareto optimal bound for the class of watermarks defined in Equation (2). The OPT watermark attains this bound, thereby validating these assumptions. We stress that this is non-trivial, as the optimal bound is computed based on properties of the original LLM only and without reference to any watermark. In addition, we note that the optimal bound can be used for tuning of the hyperparameter γ, which determines the fraction of green-list tokens (see Appendix C for details).

Also note that, strictly speaking, solutions with $\beta < 0$ are not Pareto optimal, because $\beta = 0$ is better. This means that, as long as $\beta < 0$, increasing test quality also increases text quality. In Section 3 this was associated with the greediness of the watermarking strategy. Finally, we also ran the experiments for the other optimal watermark OPT$'$, defined in Equation (8). We found that OPT and OPT$'$ perform rather similarly. See Appendix E for the results of these experiments.

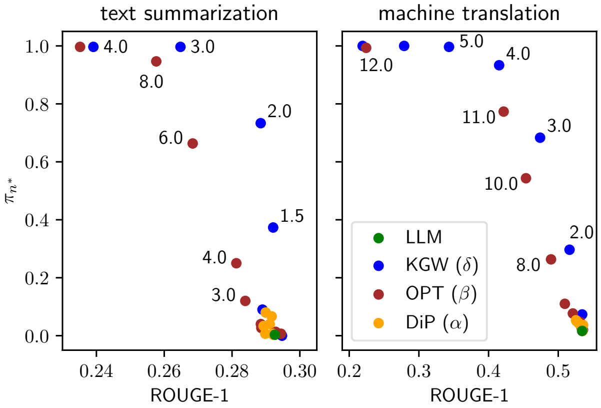

Figure 3. The power of the test versus the ROUGE-1 score, a metric for text quality, for the TS and MT tasks. For the DiP watermark the same parameter values as in Figures 1 and 2 are used. The ITS watermark has much lower ROUGE-1 scores, as it is incompatible with beam search.

4.2. Performance on other Text Quality Metrics

The performance of LLMs is commonly assessed with a variety of metrics, often depending on the specific task and type of model under consideration. Examples are perplexity, ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002). It is well known that these metrics measure different aspects of an LLM and are not necessarily in agreement with each other. This means that it is not guaranteed that the OPT watermark, which has been optimized for a low perplexity, also outperforms the other watermarks when text quality is measured differently.

To illustrate this, we perform two experiments (inspired by Hu et al. (2024)) consisting of typical tasks of LLMs: text summarization (TS) and machine translation (MT). Instead of log-perplexity, we now measure text quality in terms of ROUGE-1, which quantifies similarity between two texts in terms of the overlap of uni-grams. ROUGE scores are standard metrics to assess LLMs for summarization and translation.

For TS we use the BART-large model (Liu et al., 2020) and apply this to a randomly selected subset of the CNN-DM (test) dataset (Hermann et al., 2015). For MT we use the WMT 2016 dataset and use the Multilingual BART model (Liu et al., 2020) to translate from English to Romanian. Both models were fine-tuned and the sample size of both experiments is 300.

Figure 3 shows the test-text trade-off for both experiments and the different watermarks under consideration. Preferable is a large power of the test, corresponding to a high identifiability, and a large ROUGE-1 score close to the original LLM, corresponding to a high text quality and a stealthy watermark. We see that for TS as well as MT our OPT watermark is outperformed by the KGW watermark. This shows that watermarks, optimized for one metric, are not necessarily optimal or superior in terms of another metric. Optimized watermarks should therefore be used with caution.

Note that the ITS watermark is missing in the results. The ROUGE-1 score of ITS was very poor, as the deterministic sampling strategy is incompatible with beam search, which is standard for these types of experiments. Also note that the DiP watermark shows its distortion-free property by having a ROUGE score close to the original LLM, but performs poorly in terms of identifiability. This is in accordance with what was observed in Wu et al. (2023).

4.3. Biasedness and Dependence within Watermarks

This section investigates the validity of the assumptions that are used in the derivation of the Pareto optimal solutions to optimization problems like Equation (7), which involves the power of the test. The derivation uses that N_g is binomially distributed, which requires two additional assumptions. The first one is unbiasedness, i.e., $\mathrm{E}[\Gamma_t] = \gamma$, which is discussed in Appendix D. The second assumption is that the events that a token is a green-list token are independent and identically distributed for different tokens in the same sequence of watermarked text. The latter can be stated as that $\tilde{\mathrm{E}}[\Delta(p_t, \mathcal{G}_t)]$ must be the same for all t. This is a mild assumption, as this is an unconditional expectation value and therefore does not depend on the prefix of token t.

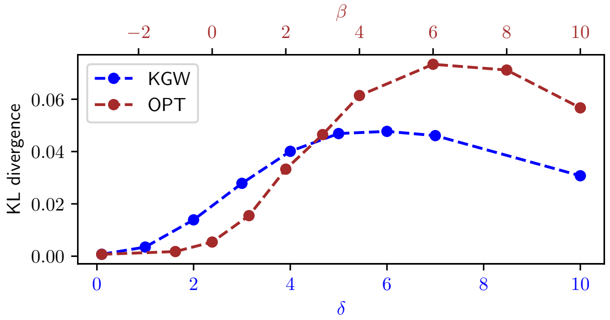

Figure 4. The Kullback-Leibler divergence between a binomial distribution for N_g, based on the empirical fraction of green-list tokens, and the empirical distribution of N_g, as a function of the strength (δ and β) of the respective watermarks.

However, it turns out that independence is violated and that the dependence becomes stronger for stronger watermarks. Figure 4 shows the Kullback-Leibler divergence between an exact binomial distribution and the empirical distribution of N_g. The KL-divergence increases with the strength of the watermarks, indicating a stronger (stochastic) dependence. For very strong watermarks the KL-divergence decreases again, as there is less space to deviate when almost all tokens are green-list tokens. Additional evidence of the breakdown of the binomial assumption can be found in Appendix D. It can be explained by the fact that not all texts allow for the same semantic freedom. Hence, the events of tokens being a green-list token are positively correlated within a sequence.

It is important to realize that Pareto optimality of the OPT watermark and the Pareto optimal bound both depend on this assumption. In our experiments its breakdown seems of little consequence, indicated by the fact that the OPT watermark outperforms other watermarks and that it coincides with the Pareto optimal bound in Figures 1 and 2. But the impact might be bigger in situations with less semantic freedom, for example in code generation (Wang et al., 2024; Lee et al., 2023).

4.4. The Impact of Watermarks on Text Diversity

Watermarks like KGW and OPT are expected to decrease the n-gram diversity, which is the fraction of unique n-grams within the n-grams of a text corpus (Li et al., 2016). The reasoning is that at position t, given the prefix $V_{:t}$, the watermark (potentially) gives preference to green-list tokens, which form a randomly selected subset of the total vocabulary $\mathcal{V}$. Also recall our interpretation of the OPT watermark with $\beta \leq 0$ as greedier than the original LLM. Figure 5 indeed shows that bi- and tri-gram diversity decrease as a consequence of the watermarks, an effect that is relatively large for the OPT watermark.

The latter is a potential disadvantage of OPT compared to other watermarks. However, n-gram diversity is not the same as semantic diversity or how humans perceive diversity. In fact, there is evidence that n-gram diversity is a poor metric for diversity judgements from humans (Tevet & Berant, 2021). It requires further research to find out whether this concern about OPT watermarks is warranted. Note that, although distortion-free, DiP also shows a decrease in bi- and tri-gram diversity. It is unbiased at the level of individual tokens, but it is not invisible (Liu et al., 2024b) when considering pairs or triplets of consecutive tokens.

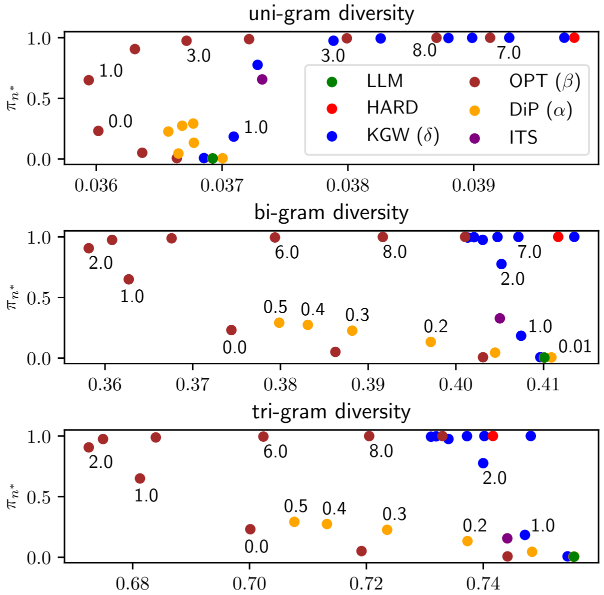

Figure 5. Uni-, bi- and tri-gram diversities for different watermarks. An n-gram diversity is the fraction of unique n-grams within the total set of n-grams of a text corpus. The results are based on the same data as Figures 1 and 2.

Another interesting feature of Figure 5 is that, when the KGW and OPT watermarks approach the hard watermark (large δ or β, respectively), the n-gram diversity starts to increase again. A possible explanation is that for these ‘harder’ watermarks a more diverse set of rather unlikely green-list tokens are sampled at the expense of more standard red-list tokens. This also explains why a hard watermark has a larger uni-gram diversity than the original LLM.

5. Related Work

Watermarking LLM-generated texts is an example of steganography, the practice of representing information (e.g., a watermark) in other information (e.g., a text). The idea of watermarking machine-learning generated text has been around for some time (Venugopal et al., 2011; Ziegler et al., 2019) and has different realizations.

A first option is to impose a distribution shift on the LLM that is tractable for a detector, with Kirchenbauer et al. (2023) as the most prominent example. Since then, mostly in view of mitigating between the four criteria for watermarks outlined in Section 1, a myriad of alternatives has been proposed (Zhao et al., 2024; Wang et al., 2024; Fang et al., 2023; Lee et al., 2023; Liu et al., 2024a; Fu et al., 2024; Liu et al., 2024b; Chen et al., 2024). A notable subclass of examples, including DiP, are unbiased or distortion-free watermarks. This means that the distribution shift of the watermark is unbiased when averaged over the pseudo-random aspect of the watermark (Hu et al., 2024; Wu et al., 2023).

A second option, which includes ITS, is to instill a watermark into the pseudo-random sampling from the LLM (Christ et al., 2023; Kuditipudi et al., 2024). This also has the advantage that it can be made distortion-free. However, examples come also with disadvantages of low robustness or inefficiency (Wu et al., 2023).

A third option is a watermark based on an additional ML-model (Abdelnabi & Fritz, 2021; Qiang et al., 2023; Yoo et al., 2023; Yang et al., 2023; Munyer et al., 2024). The watermark is then imposed by making alterations to a text that is already generated by the LLM. This requires extra training and due to the extra flexibility ensuring the quality of the watermark can be difficult. Finally, one could instill a watermark into the weights of the LLM by adjusting the training procedure (Li et al., 2023a; Gu et al., 2024).

Another branch of this developing field is the analysis of watermarks for LLMs. This includes the design of benchmark tasks and metrics to test the quality of watermarks (Piet et al., 2023), some with a special focus on text quality (Tang et al., 2023; Tu et al., 2023; Ajith et al., 2023) or robustness (Krishna et al., 2023; Sadasivan et al., 2023; Shi et al., 2024; Kirchenbauer et al., 2024; Zhang et al., 2023).

To conclude, watermarking is not the only option to distinguish LLM-generated texts from human-generated texts. An alternative is to train a binary classifier to detect LLM-generated texts (see, e.g., Mitchell et al. (2023)). With the rapid improvement of LLMs, this has become increasingly difficult (Gambini et al., 2022). Another option is to let the vendor of the LLM keep a copy of all generated output and provide an API that compares a text with this database of outputs (Krishna et al., 2023).

6. Conclusion and Discussion

It was posited in Section 1 that the contribution of this paper is twofold. It introduces a new watermark, OPT, corresponding to the Pareto optimal solutions of the multi-objective optimization problem into which the test-text trade-off was translated. Since the watermarks over which we optimize are generally considered robust and efficient, we believe that OPT has an excellent standing with respect to the four criteria for good watermarks for LLMs: identifiability, stealthiness, robustness and efficiency. This is notwithstanding the provisions that were made about Pareto optimal watermarks in Sections 4.2, 4.3 and 4.4.

But this paper should also be read as the introduction of a systematic approach to optimizing the test-text trade-off for watermarks of LLMs. The chosen translation of the trade-off into an optimization problem is not unique, as it depends on how you quantify test and text quality. And also the class of watermarks over which we optimize, defined in Equation (2), is not unique. It was chosen to be based on a green-red split of the vocabulary, such that it is a generalization of the original watermark of Kirchenbauer et al. (2023). And its form was chosen to be so-called *minimal*, i.e., all green-list probabilities are rescaled by the same factor and the same holds for all red-list tokens.

It is conceivable that different choices regarding the above lead, after optimization, to more preferable watermarks. One option is to remove some of the implicit restrictions that are imposed by Equation (2). The shift function $\Delta(p_t, \mathcal{G}_t)$ that determines $\tilde{\mathrm{P}}[V_t|V_{:t}]$ is the same for each $V_t \in \mathcal{V}$, but this does not have to be the case. Also note that the shift function is determined by properties of token t alone. One could try to make it dependent on subsequent tokens; a choice for a red token at position t could enable a string of green tokens in what follows. Another possibility is to keep track of the number of green-list tokens already generated and use this to adjust the shift function.

Acknowledgements

The author would like to thank Cees Diks and Floris Holstege for useful discussions and feedback on the manuscript.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. The general topic of this paper, watermarks for large-language models, is specifically aimed at mitigating potential societal risks of Machine Learning, some of which were outlined in Section 1. This paper aims to contribute to this.

References

Aaronson, S. My AI Safety Lecture for UT Effective Altruism, 2022. URL https://scottaaronson.blog/?p=6823.

Abdelnabi, S. and Fritz, M. Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding. *2021 IEEE Symposium on Security and Privacy (SP)*, 00:121–140, 2021. doi: 10.1109/sp40001.2021.00083.

Ajith, A., Singh, S., and Pruthi, D. Performance Tradeoffs of Watermarking Large Language Models. *arXiv*, 2023. doi: 10.48550/arxiv.2311.09816. URL https://arxiv.org/abs/2311.09816.

Bartz, D. and Hu, K. OpenAI, Google, others pledge to watermark AI content for safety, White House says, 2023. URL https://www.reuters.com/technology/openai-google-others-pledge-watermark-ai-content-safety-white-house-2023-07-21/.

Chen, L., Bian, Y., Deng, Y., Cai, D., Li, S., Zhao, P., and Wong, K.-f. WatME: Towards Lossless Watermarking Through Lexical Redundancy. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024. URL https://openreview.net/forum?id=f4SLQEAePH.

Christ, M., Gunn, S., and Zamir, O. Undetectable Watermarks for Language Models. *arXiv*, 2023. doi: 10.48550/arxiv.2306.09194. URL https://arxiv.org/abs/2306.09194.

Fang, J., Tan, Z., and Shi, X. COSYWA: Enhancing Semantic Integrity in Watermarking Natural Language Generation. *Lecture Notes in Computer Science*, pp. 708–720, 2023. ISSN 0302-9743. doi: 10.1007/978-3-031-44693-1_55.

Fu, Y., Xiong, D., and Dong, Y. Watermarking Conditional Text Generation for AI Detection: Unveiling Challenges and a Semantic-Aware Watermark Remedy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18003–18011, 2024. doi: 10.1609/aaai.v38i16.29756. URL https://ojs.aaai.org/index.php/AAAI/article/view/29756.

Gambini, M., Fagni, T., Falchi, F., and Tesconi, M. On pushing DeepFake Tweet Detection capabilities to the limits. *14th ACM Web Science Conference 2022*, pp. 154–163, 2022. doi: 10.1145/3501247.3531560.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., and Sedova, K. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *arXiv*, 2023. doi: 10.48550/arxiv.2301.04246. URL https://arxiv.org/abs/2301.04246.

Gu, C., Li, X. L., Liang, P., and Hashimoto, T. On the Learnability of Watermarks for Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9k0krNzvlV.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28, 2015.

Hu, Z., Chen, L., Wu, X., Wu, Y., Zhang, H., and Huang, H. Unbiased Watermark for Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uWVC5FVidc.

Jiang, Z., Zhang, J., and Gong, N. Z. Evading Watermark based Detection of AI-Generated Content. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1168–1181, 2023. doi: 10.1145/3576915.3623189.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A Watermark for Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *arXiv*, pp. 17061—17084. PMLR, 2023. doi: 10.48550/arxiv.2301.10226. URL https://arxiv.org/abs/2301.10226.

Kirchenbauer, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., Fernando, K., Saha, A., Goldblum, M., and Goldstein, T. On the Reliability of Watermarks for Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., and Iyyer, M. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Kuditipudi, R., Thickstun, J., Hashimoto, T., and Liang, P. Robust Distortion-free Watermarks for Language Models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Lee, T., Hong, S., Ahn, J., Hong, I., Lee, H., Yun, S., Shin, J., and Kim, G. Who Wrote this Code? Watermarking for Code Generation. *arXiv*, 2023. doi: 10.48550/arxiv.2305.15060. URL https://arxiv.org/abs/2305.15060.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. A Diversity-Promoting Objective Function for Neural Conversation Models. *Proceedings of the 2016 Conference of* *the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, 2016. doi: 10.18653/v1/n16-1014.

Li, L., Jiang, B., Wang, P., Ren, K., Yan, H., and Qiu, X. Watermarking LLMs with Weight Quantization. *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3368–3378, 2023a. doi: 10.18653/v1/2023.findings-emnlp.220.

Li, Y., Wang, Y., Shi, Z., and Hsieh, C.-J. Improving the Generation Quality of Watermarked Large Language Models via Word Importance Scoring. *arXiv*, 2023b. doi: 10.48550/arxiv.2311.09668. URL https://arxiv.org/abs/2311.09668.

Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013.

Liu, A., Pan, L., Hu, X., Li, S., Wen, L., King, I., and Yu, P. S. An Unforgeable Publicly Verifiable Watermark for Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024a.

Liu, A., Pan, L., Hu, X., Meng, S., and Wen, L. A Semantic Invariant Robust Watermark for Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024b.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020. doi: 10.1162/tacl_a_00343.

Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., O'Connor, K., Li, R., Peng, P.-C., Bright, T. J., Tatonetti, N., Won, K. J., Gonzalez-Hernandez, G., and Moore, J. H. ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining*, 16(1):20, 2023. ISSN 1756-0381. doi: 10.1186/s13040-023-00339-9.

Milano, S., McGrane, J. A., and Leonelli, S. Large language models challenge the future of higher education. *Nature Machine Intelligence*, 5(4):333–334, 2023. doi: 10.1038/s42256-023-00644-2.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Munyer, T., Tanvir, A. A., Das, A., Zhong, X., and ZHONG, X. DeepTextMark: A Deep Learning-Driven Text Watermarking Approach for Identifying Large Language Model Generated Text. *IEEE Access*, 12:40508–40520, 2024. ISSN 2169-3536. doi: 10.1109/access.2024.3376693.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, pp. 311–318, 2002. doi: 10.3115/1073083.1073135.

Piet, J., Sitawarin, C., Fang, V., Mu, N., and Wagner, D. Mark My Words: Analyzing and Evaluating Language Model Watermarks. *arXiv*, 2023. doi: 10.48550/arxiv.2312.00273. URL https://arxiv.org/abs/2312.00273.

Qiang, J., Zhu, S., Li, Y., Zhu, Y., Yuan, Y., and Wu, X. Natural language watermarking via paraphraser-based lexical substitution. *Artificial Intelligence*, 317:103859, 2023. ISSN 0004-3702. doi: 10.1016/j.artint.2023.103859.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21, 2020. ISSN 1532-4435. doi: 10.5555/3455716.3455856.

Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A., and Sauerland, U. Risks and Benefits of Large Language Models for the Environment. *Environmental Science & Technology*, 57(9):3464–3466, 2023. ISSN 0013-936X. doi: 10.1021/acs.est.3c01106.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. Can AI-Generated Text be Reliably Detected? *arXiv*, 2023. doi: 10.48550/arxiv.2303.11156. URL https://arxiv.org/abs/2303.11156.

Shi, Z., Wang, Y., Yin, F., Chen, X., Chang, K.-W., and Hsieh, C.-J. Red Teaming Language Model Detectors with Language Models. *Transactions of the Association for Computational Linguistics*, 12:174–189, 2024. doi: 10.1162/tacl_a_00639.

Tang, L., Uberti, G., and Shlomi, T. Baselines for Identifying Watermarked Large Language Models. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023.

Tevet, G. and Berant, J. Evaluating the Evaluation of Diversity in Natural Language Generation. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 326–346, 2021. doi: 10.18653/v1/2021.eacl-main.25.

Tu, S., Sun, Y., Bai, Y., Yu, J., Hou, L., and Li, J. WaterBench: Towards Holistic Evaluation of Watermarks for Large Language Models. *arXiv*, 2023. doi: 10.48550/arxiv.2311.07138. URL https://arxiv.org/abs/2311.07138.

Venugopal, A., Uszkoreit, J., Talbot, D., Och, F., and Ganitkevitch, J. Watermarking the Outputs of Structured Prediction with an application in Statistical Machine Translation. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1363—1372, 2011. URL https://aclanthology.org/D11-1126.

Vincent, J. AI-generated answers temporarily banned on coding Q&A site Stack Overflow, 2022. URL https://www.theverge.com/2022/12/5/23493932/chatgpt-ai-generated-answers-temporarily-banned-stack-overflow-llms-dangers.

Wang, L., Yang, W., Chen, D., Zhou, H., Lin, Y., Meng, F., Zhou, J., and Sun, X. Towards Codable Text Watermarking for Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. v., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

Wu, Y., Hu, Z., Zhang, H., and Huang, H. DiPmark: A Stealthy, Efficient and Resilient Watermark for Large Language Models. *arXiv*, 2023. doi: 10.48550/arxiv.2310.07710. URL https://arxiv.org/abs/2310.07710.

Yang, X., Chen, K., Zhang, W., Liu, C., Qi, Y., Zhang, J., Fang, H., and Yu, N. Watermarking Text Generated by Black-Box Language Models. *arXiv*, 2023. doi: 10.48550/arxiv.2305.08883. URL https://arxiv.org/abs/2305.08883.

Yoo, K., Ahn, W., Jang, J., and Kwak, N. Robust Multi-bit Natural Language Watermarking through Invariant Features. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2092–2115, 2023. doi: 10.18653/v1/2023.acl-long.117.

Zhang, H., Edelman, B. L., Francati, D., Venturi, D., Ateniese, G., and Barak, B. Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models. *arXiv*, 2023. doi: 10.48550/arxiv.2311.04378. URL https://arxiv.org/abs/2311.04378.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. OPT: Open Pre-trained Transformer Language Models. *arXiv*, 2022. doi: 10.48550/arxiv.2205.01068. URL https://arxiv.org/abs/2205.01068.

Zhao, X., Ananth, P., Li, L., and Wang, Y.-X. Provable Robust Watermarking for AI-Generated Text. In *The Twelfth International Conference on Learning Representations*, 2024.

Ziegler, Z., Deng, Y., and Rush, A. Neural Linguistic Steganography. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1210–1215, 2019. doi: 10.18653/v1/d19-1115.

A. Optimization objectives and their solutions

A.1. Measures for test quality

One measure for test quality is N_g, the number of green-list tokens in a watermarked sequence of length T. A strong test, i.e. a test with a large identifiability, corresponds to a large number of green-list tokens. If we define binary random variables Y_t such that

$$Y_t = \begin{cases} 1 & \text{if } V_t \in \mathcal{G}_t, \\ 0 & \text{if } V_t \notin \mathcal{G}_t, \end{cases} \tag{9}$$

then $N_g = Y_1 + Y_2 + \ldots + Y_T$. The conditional random variables $Y_t | V_{:t}$ are Bernoulli distributed,

$$Y_t | V_{:t} \sim \text{BIN}(1, \Gamma_t) \tag{10}$$

in the absence of a watermark and

$$Y_t | V_{:t} \sim \text{BIN}(1, \Gamma_t + \Delta(p_t, \mathcal{G}_t)) \tag{11}$$

in the presence of a watermark. The shift, due to a watermark defined by Equation (2), in the expected number of green-list tokens is given by Equation (3) and can be derived as follows:

$$\begin{aligned}
\tilde{\text{E}}[N_g] &= \sum_{t=1}^{T} \tilde{\text{E}}\left[\tilde{\text{E}}[Y_t | V_{:t}]\right] && \text{(12a)} \\
&= \sum_{t=1}^{T} \tilde{\text{E}}[\Gamma_t + \Delta(p_t, \mathcal{G}_t)] && \text{(12b)} \\
&= \sum_{t=1}^{T} \left\{ \tilde{\text{E}}[\Gamma_t] + \tilde{\text{E}}[\Delta(p_t, \mathcal{G}_t)] \right\} && \text{(12c)} \\
&= \sum_{t=1}^{T} \left\{ \text{E}[\Gamma_t] + \tilde{\text{E}}[\Delta(p_t, \mathcal{G}_t)] \right\} && \text{(12d)} \\
&= \text{E}[N_g] + \sum_{t=1}^{T} \tilde{\text{E}}[\Delta(p_t, \mathcal{G}_t)] && \text{(12e)}
\end{aligned}$$

where in (12d) we used the assumption that expectations are unaffected by watermark-induced changes in the distribution of the prefix $V_{:t}$. Explicitly, the assumption that $\tilde{\text{E}}[\Gamma_t] = \text{E}[\Gamma_t]$ means

$$\sum_{v_{:t} \in \mathcal{V}^*} \text{P}[V_t \in \mathcal{G}_t | V_{:t} = v_{:t}] \tilde{\text{P}}[V_{:t} = v_{:t}] = \sum_{v_{:t} \in \mathcal{V}^*} \text{P}[V_t \in \mathcal{G}_t | V_{:t} = v_{:t}] \text{P}[V_{:t} = v_{:t}] \,, \tag{13}$$

where $\mathcal{V}^*$ is the space of all possible prefixes of arbitrary length. It should be emphasized that this is an approximate assumption, based on the idea that watermark-induced changes in the distribution of the prefix $V_{:t}$ are (approximately) averaged out in the sum over $\mathcal{V}^*$. We also emphasize that this assumption does not mean that the marginal distribution of the prefix $V_{:t}$ is unaffected by the watermark, which would be a much stronger assumption.

Another measure for test quality is the power of the test, $\pi_{n^*} = \tilde{\text{P}}[N_g \geq n^*]$. If we assume that $\tilde{\text{E}}[\Gamma_t] = \text{E}[\Gamma_t] = \gamma$, that the Y_t are identically distributed and that they are (stochastically) independent, then

$$N_g \sim \text{BIN}(T, \gamma + \tilde{\text{E}}[\Delta(p_t, \mathcal{G}_t)]), \tag{14}$$

because $\tilde{\text{E}}[Y_t] = \tilde{\text{E}}\left[\tilde{\text{E}}[Y_t | V_{:t}]\right] = \tilde{\text{E}}[\Gamma_t + \Delta(p_t, \mathcal{G}_t)]$. This implies

$$\pi_{n^*} = \sum_{n=n^*}^{T} \binom{T}{n} \left(\gamma + \tilde{\text{E}}[\Delta(p_t, \mathcal{G}_t)]\right)^n \left(1 - \gamma - \tilde{\text{E}}[\Delta(p_t, \mathcal{G}_t)]\right)^{T-n} . \tag{15}$$

In other words, the watermark determines the power of the test only through $\tilde{\text{E}}[\Delta(p_t, \mathcal{G}_t)]$ and does this in a monotonically increasing way. Hence, maximizing the power of the test over a class of watermarks is equivalent to maximizing N_g.

A.2. Measures for text quality

One measure for text quality is the log-perplexity, defined in Section 3. A large log-perplexity is interpreted as a low text quality. The shift, due to a watermark defined by Equation (2), in the expected log-perplexity is given by Equation (4) and can be derived as follows:

$$\begin{aligned}
\tilde{\mathrm{E}}[\log \mathrm{PPL}] &= -\frac{1}{T}\sum_{t=1}^{T} \tilde{\mathrm{E}}[\log \mathrm{P}[V_t|V_{:t}]] && \text{(16a)}\\
&= -\frac{1}{T}\sum_{t=1}^{T} \tilde{\mathrm{E}}\left[\sum_{v\in\mathcal{V}} \tilde{\mathrm{P}}[V_t = v|V_{:t}] \log \mathrm{P}[V_t = v|V_{:t}]\right] && \text{(16b)}\\
&= -\frac{1}{T}\sum_{t=1}^{T} \tilde{\mathrm{E}}[\mathrm{E}[\log \mathrm{P}[V_t|V_{:t}] \mid V_{:t}]] && \text{(16c)}\\
&\quad -\frac{1}{T}\sum_{t=1}^{T} \tilde{\mathrm{E}}\left[\sum_{v\in\mathcal{V}} \Delta(p_t,\mathcal{G}_t)\left\{\frac{\mathrm{I}_{\mathcal{G}_t}(v)}{\Gamma_t} - \frac{1-\mathrm{I}_{\mathcal{G}_t}(v)}{1-\Gamma_t}\right\} \mathrm{P}[V_t = v|V_{:t}] \log \mathrm{P}[V_t = v|V_{:t}]\right] && \text{(16d)}\\
&= -\frac{1}{T}\sum_{t=1}^{T} \tilde{\mathrm{E}}[\mathrm{E}[\log \mathrm{P}[V_t|V_{:t}] \mid V_{:t}]] + \frac{1}{T}\sum_{t=1}^{T} \tilde{\mathrm{E}}[\Delta(p_t,\mathcal{G}_t)B(p_t,\mathcal{G}_t)] && \text{(16e)}\\
&= -\frac{1}{T}\sum_{t=1}^{T} \mathrm{E}[\log \mathrm{P}[V_t|V_{:t}]] + \frac{1}{T}\sum_{t=1}^{T} \tilde{\mathrm{E}}[\Delta(p_t,\mathcal{G}_t)B(p_t,\mathcal{G}_t)] && \text{(16f)}\\
&= \mathrm{E}[\log \mathrm{PPL}] + \frac{1}{T}\sum_{t=1}^{T} \tilde{\mathrm{E}}[\Delta(p_t,\mathcal{G}_t)B(p_t,\mathcal{G}_t)]\,. && \text{(16g)}
\end{aligned}$$

In (16e) we used the definition of $B(p_t, \mathcal{G}_t)$, see Section 3. In (16f) we used a similar approximate assumption as in (12d). See (13) for a discussion.

A.3. Derivation of Pareto optimal solutions

The multi-objective optimization problem of Equation (5) can be written as

$$\max_{\Delta\in\Upsilon} \tilde{\mathrm{E}}[\Delta(p_t,\mathcal{G}_t)] \qquad \text{and} \qquad \min_{\Delta\in\Upsilon} \tilde{\mathrm{E}}[\Delta(p_t,\mathcal{G}_t)B(p_t,\mathcal{G}_t)]\,, \tag{17}$$

where Υ is the set of shift functions $\Delta : \Xi \times \Theta \to [0, 1]$, where Ξ is the space of all pmfs over the vocabulary $\mathcal{V}$, and Θ is the space of all subsets of size $\lfloor \gamma N \rfloor$ of the vocabulary $\mathcal{V}$, with the additional (consistency) requirement that $\Delta(p_t, \mathcal{G}_t) \leq 1 - \Gamma_t$.

The expectations $\tilde{\mathrm{E}}[\cdot]$ are taken over all possible sequences $V_1, \ldots, V_T$ generated by a watermarked LLM, and all possible prompts $V_{:1}$. However, in the definition of the optimization problem only two functions of the individual tokens play a role: Γ_t and $B(p_t, \mathcal{G}_t)$. The optimization problem can therefore be rephrased in terms of the search for a function $h(x, y)$ of a bivariate random variable (X, Y), where X plays the role of Γ_t and Y plays the role of $B(p_t, \mathcal{G}_t)$, that has support $[0, 1] \times \mathbb{R}$:

$$\max_{h(\cdot,\cdot)} \mathrm{E}_{(X,Y)}[h(X,Y)] \qquad \text{and} \qquad \min_{h(\cdot,\cdot)} \mathrm{E}_{(X,Y)}[h(X,Y)Y]\,, \tag{18}$$

where the expectations are with respect to the joint distribution of (X, Y), and where the function $h(x, y)$ obeys the constraint $0 \leq h(x, y) \leq 1 - x$. We claim that

$$h^*(x,y) = \begin{cases} 1-x & \text{if } y \leq \beta, \\ 0 & \text{if } y > \beta, \end{cases} \tag{19}$$

which corresponds to Equation (6), is Pareto optimal if $\beta \geq 0$. We begin by giving an intuitive argument for this claim, which is then followed by a more formal proof.

Intuitive argument for Pareto optimality

If $f(x, y)$ is the joint pdf of (X, Y), the optimization problem of Equation (18) can simply be rephrased as

$$\max_{\tilde{h}(\cdot,\cdot)} \int \left[\int \tilde{h}(x, y) \, \mathrm{d}y \right] \mathrm{d}x \qquad \text{and} \qquad \min_{\tilde{h}(\cdot,\cdot)} \int \left[\int \tilde{h}(x, y) y \, \mathrm{d}y \right] \mathrm{d}x, \tag{20}$$

where $\tilde{h}(x, y) = h(x, y) f(x, y)$ and the constraint is now $0 \leq \tilde{h}(x, y) \leq (1 - x) f(x, y)$. We see that for each value of x, this is a separate multi-objective optimization problem. Given a value for x, we want

$$\max_{\tilde{h}(x,\cdot)} \int \tilde{h}(x, y) \, \mathrm{d}y \qquad \text{and} \qquad \min_{\tilde{h}(x,\cdot)} \int \tilde{h}(x, y) y \, \mathrm{d}y. \tag{21}$$

In words, we are looking for a function $\tilde{h}(x, \cdot)$ of only a single variable y. We want its area under the curve to be maximized. At the same time, we want the volume of variable height y above this area to be minimized.

Suppose we want the area under the curve to be of a certain value, how do we approach this problem? The solution is to start making the function $\tilde{h}(x, \cdot)$ maximal for the smallest possible values of y, because for those values the contribution to the volume is the smallest. You continue doing this for larger and larger values of y, until you have reached the desired area. This will minimize the volume. If, on the other hand, you wish the volume to be of a certain value, you take the same approach until you have reached that volume. This will maximize the area under the curve.

To make the optimization problem of Equation (21) more tangible, imagine the following analogy. You are standing at location A next to a pile of stones of variable weights. If someone asks you to bring 10 stones to location B, you choose to bring the 10 lightest stones to make your life easy. If someone asks you to bring as many stones as possible to location B, but with a total weight of 5kg. Then you again choose to bring the lightest stones until you have reached a total 5kg. In this analogy the stones are selected by the function $\tilde{h}(x, \cdot)$, which plays the role of an indicator function, whereas the weight of the individual stones is the variable y.

Formal proof of Pareto optimality

We continue with a more formal proof of the Pareto optimality of the function $h^*(x, y)$ defined in Equation (19). We do this by showing that a function $h'(x, y)$ that obeys the same constraints and that improves the first objective in Equation (18), necessarily does worse for the second objective in Equation (18).

Hence, assume that $\mathrm{E}_{(X,Y)}[h'(X, Y)] > \mathrm{E}_{(X,Y)}[h^*(X, Y)]$. Suppose $h'(x, y) \neq h^*(x, y)$ if $y \in A_1 \cup A_2$, where $A_1 \subset (-\infty, \beta]$ and $A_2 \in (\beta, \infty)$. This means that $h'(x, y) < 1 - x$ if $y \in A_1 \subset (-\infty, \beta]$ and $h'(x, y) > 0$ if $y \in A_2 \subset (\beta, \infty)$. The initial assumption then translates into

$$\int_{A_1} \int_0^1 \left[(1 - x) - h'(x, y) \right] f(x, y) \, \mathrm{d}x \, \mathrm{d}y < \int_{A_2} \int_0^1 h'(x, y) f(x, y) \, \mathrm{d}x \, \mathrm{d}y, \tag{22}$$

where $f(x, y)$ is the joint pdf of (X, Y). Note that A_1 can be empty. Let's now focus on the second objective, which is to minimize (assuming $\beta > 0$)

$$
\begin{aligned}
\mathrm{E}[h'(x,y)y] &= \int_{-\infty}^{\infty}\int_0^1 h'(x,y)yf(x,y)\,\mathrm{d}x\,\mathrm{d}y && \text{(23a)}\\
&= \int_{-\infty}^{\infty}\int_0^1 h^*(x,y)yf(x,y)\,\mathrm{d}x\,\mathrm{d}y && \text{(23b)}\\
&\quad + \int_{A_2}\int_0^1 h'(x,y)yf(x,y)\,\mathrm{d}x\,\mathrm{d}y - \int_{A_1}\int_0^1 \left[(1-x) - h'(x,y)\right]yf(x,y)\,\mathrm{d}x\,\mathrm{d}y && \text{(23c)}\\
&> \int_{-\infty}^{\infty}\int_0^1 h^*(x,y)yf(x,y)\,\mathrm{d}x\,\mathrm{d}y && \text{(23d)}\\
&\quad + \beta\left[\int_{A_2}\int_0^1 h'(x,y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y - \int_{A_1}\int_0^1 \left[(1-x) - h'(x,y)\right]f(x,y)\,\mathrm{d}x\,\mathrm{d}y\right] && \text{(23e)}\\
&> \int_{-\infty}^{\infty}\int_0^1 h^*(x,y)yf(x,y)\,\mathrm{d}x\,\mathrm{d}y && \text{(23f)}\\
&= \mathrm{E}[h^*(x,y)y] && \text{(23g)}
\end{aligned}
$$

where in (23f) we used that $\mathrm{E}_{(X,Y)}[h'(X,Y)] > \mathrm{E}_{(X,Y)}[h^*(X,Y)]$ and (23d) is a strict inequality because $\mathrm{E}_{(X,Y)}[h'(X,Y)] > \mathrm{E}_{(X,Y)}[h^*(X,Y)]$ is a strict inequality. Note that the second inequality is an equality in the special case $\beta = 0$. This derivation shows that the second objective is worse off. One can make a similar argument for improving on the second objective in Equation (18) and showing that the first objective will then necessarily become worse.

As already explained in Appendix (A.1), the multi-objective optimization problem defined by Equation (7) has the same Pareto optimal solutions.

The watermarks OPT', defined in Equation (8), can be shown to be Pareto optimal in a similar fashion. The only difference is that $-\log \mathrm{P}[V_t = v|V_{:t}]$ must be replaced by $[-\log \mathrm{P}[V_t = v|V_{:t}]]^2$.

B. Experimental details

All experiments have been implemented by means of the Huggingface library (Wolf et al., 2020).

B.1. Details of the experiments described in Sections 4.1, 4.3, 4.4

Prompts are created by randomly selecting (news) texts from the C4 dataset (Raffel et al., 2020). Only texts of at least 250 tokens are taken into account. If a text has at most 400 tokens, the final 200 tokens are removed and the remainder forms the prompt. If a text has more than 400 tokens, the first 200 tokens are used as prompt and the rest is discarded. All prompts have a length between 50 and 200 tokens.

Sampling from the LLM takes place with a temperature of 1.0. In order to generate sequences of a fixed length T, the EOS token is suppressed.

B.1.1. Estimation of the log-perplexity

Suppose $v_1, v_2, \ldots, v_T$ is an actual realization of the random variables $V_1, V_2, \ldots, V_T$. The log-perplexity of this sequence is

$$-\frac{1}{T}\sum_{t=1}^{T} \log \mathrm{P}[V_t = v_t|V_{:t} = v_{:t}]. \tag{24}$$

In order to estimate $\tilde{\mathrm{E}}[\log \mathrm{PPL}]$, the above formula could be used. However, in order to reduce estimation noise, we used

$$-\frac{1}{T}\sum_{t=1}^{T}\sum_{v\in\mathcal{V}} \tilde{\mathrm{P}}[V_t = v|V_{:t} = v_{:t}] \log \mathrm{P}[V_t = v|V_{:t} = v_{:t}]. \tag{25}$$

B.2. Details of the experiments described in Section 4.2

For the text summarization (TS) and machine translation (MT) tasks, we have used the default setup. Samples of size 300 were randomly selected from the respective test sets. We used the default sampling strategy: beam search with 4 beams.

B.3. A note about the implementation of other watermarks

The vanilla implementation of the ITS watermark (Kuditipudi et al., 2024) lead to a much higher log-perplexity that one would expect from a distortion-free watermark. Tokens with extremely small probabilities are sampled with unrealistically high frequencies. Even tokens with a probability of strictly zero were sometimes sampled. We believe that the underlying cause is numerical inprecision when selecting the index that makes the cumulative distribution function pass a certain threshold. To address this problem, we forbid sampling from tokens with probability below 10e-6.

Furthermore, for the ITS watermark we do not use Levenshtein distance, because their purpose is to robustify the watermark against insertions and deletions and we did not consider this.

C. Hyperparameter tuning

Watermarks defined by Equation (2) have a hyperparameter γ, the fraction of the vocabulary $\mathcal{V}$ that is on the green list. A large γ means that for each token relatively many green-list options are available. This makes the expected deterioration of text quality relatively small. However, when γ is large a powerful test also requires relatively many green tokens in a sequence of length T. For small γ the trade-off is vice versa. This raises the question of an optimal value of γ with respect to the test-text trade-off.

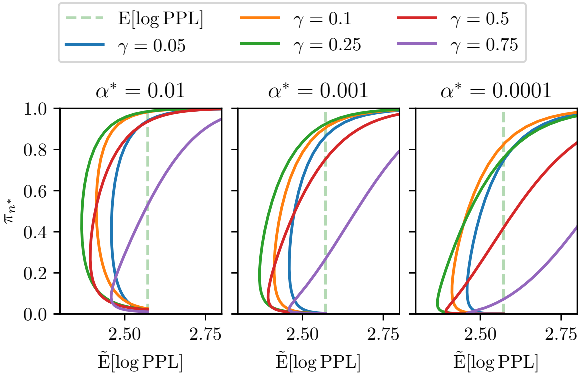

Figure 6. Pareto optimal bounds for different values of the hyperparameter γ, for tests with different false-positive rates α^*. It shows that there is no universally "best" γ.

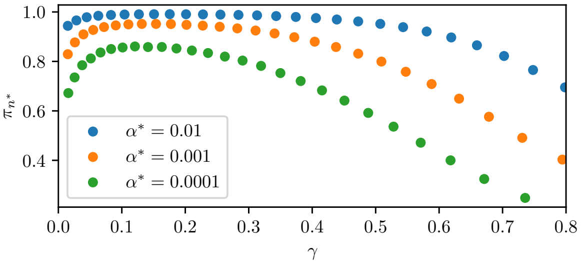

Figure 7. The power of OPT watermarks that do not affect text quality ($\tilde{\text{E}}[\log \text{PPL}] = \text{E}[\log \text{PPL}]$), as a function of the hyperparameter γ, for tests with different false-positive rates α^*. The "best" value for γ usually lies between 0.1 and 0.2.

In Figure 6 Pareto optimal bounds are plotted for different values of the hyperparameter γ. It shows that there is no universally best γ, i.e., a hyperparameter value that gives the best test-text trade-off for every possible false-positive rate α^*. One possible definition of a "best" γ is the one that, given a false-positive rate α^*, maximizes the power of the OPT watermark that does not affect text quality, i.e., $\tilde{\mathrm{E}}[\log \mathrm{PPL}] = \mathrm{E}[\log \mathrm{PPL}]$. Figure 7 shows that this "best" γ usually lies between 0.1 and 0.2, where it should be noted that the range of "near-best" hyperparameter values increases with increasing α^*. Interestingly, Kirchenbauer et al. (2023) found a "best" γ around 0.1 for their non-optimal KGW watermark. It should be emphasized that the optimal value of γ is not only dependent on how the test-text trade-off is defined, but is also a property of the LLM. A different LLM could give a different optimal value for γ.

D. Additional results on biasedness and indepedence within watermarks

As mentioned in Section 3.1, the derivation of the Pareto optimal solutions to optimization problems involving the power of the test, e.g. Equation (7), uses that N_g is binomially distributed and this requires two additional assumptions. The first one is unbiasedness of the (conditional) probability of a token being a green-list token in the absence of a watermark, $\mathrm{E}[\Gamma_t] = \gamma$. To test this, for 10,000 prompts from the C4 dataset a sequence of 30 tokens is generated without watermark. For each token Γ_t is computed for a green-red split with $\gamma = 0.25$. The sample average is $\overline{\Gamma}_t = 0.2574$. Under the null hypothesis of unbiasedness, and under the assumption of independence of the observations in the sample, this corresponds to a z-score of 16.1. This strongly suggests that $\mathrm{E}[\Gamma_t] \neq \gamma$.

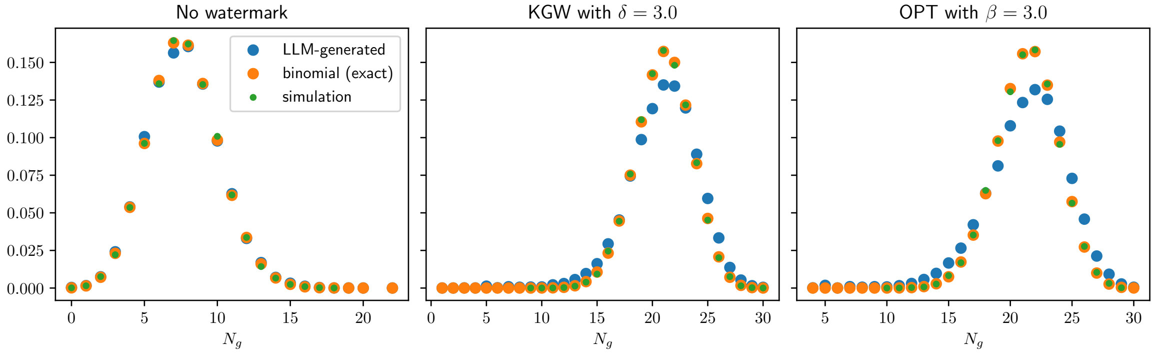

Figure 8. A comparison between the empirical distribution of N_g, the number of green-list tokens in a sequence of $T = 30$ tokens, and the exact binomial distribution, for text generated without watermark (*left panel*) and with the OPT watermark with $\beta = 4.0$ (*middle panel*). The oracle OPT (*right panel*) is the OPT watermark, but with a different, randomly selected key for green-red split generation at every generation step. Because oracle OPT and OPT differ from the exact binomial distribution similarly, we conclude the discrepancy is not because of the pseudo-random green-red split. Also included is a simulation sampled from the exact binomial distribution, of the same sample size as the LLM generated data.

This bias has its origin in the pseudo-random green-red split of the vocabulary. For each pair of subsequent tokens in a sentence, the key of the function g_γ determines whether the second token of the pair is green or red. This is the same for each occurrence of the pair. Hence, the occurrence frequencies of all possible pairs of tokens, together with the key, determine the bias of $\mathrm{E}[\Gamma_t]$ with respect to γ. We verified that a different generator g_γ, based on a different key, gives a different bias. We emphasize that it is unlikely that this bias has an effect on the Pareto optimality of the solutions presented in this paper, because also in the presence of a bias the power of the test remains a monotonic function of ΔN_g.

The second assumption, which was discussed in Section 4.3, is that the event that a token is a green-list token is (stochastically) independent for different tokens in the same sequence. It turns out that this is not the case, and that the dependence becomes stronger for stronger watermarks. Figure 8 shows additional empirical evidence that the distribution of N_g for watermarked text is generally not binomial. The heavier tails indicate a positive correlation between the events that tokens from the same list are green-list tokens. This is understandable, as different sentences can have different amounts of freedom (i.e. entropy) to insert green-list tokens.

E. Additional results

Additional descriptive statistics of the effect of different watermarks on text quality can be found in Figure 9, where the qth percentile of $-\log \mathrm{P}[V_t|V_{:t}]$ is plotted for different values of q. The relative difference in percentiles between the KGW and OPT watermarks decreases with increasing q and is virtually absent when $q = 0.99$. In other words, the OPT watermark is not better than the KGW watermark for tokens with a very large log-perplexity. This is not necessarily problematic, as the original LLM already generates these very large log-perplexities.

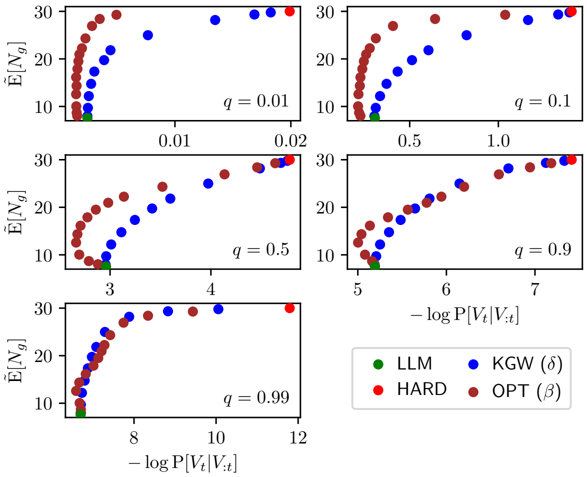

Figure 9. The qth percentile of $-\log \mathrm{P}[V_t|V_{:t}]$ is shown for different watermarks and $q = 0.01, 0.1, 0.5, 0.9$ and 0.99.

Figures 10, 11 and 12 are the same experiments as described in Section 4.1, but now including the OPT′ watermark. They were conducted with the OPT-350m model.

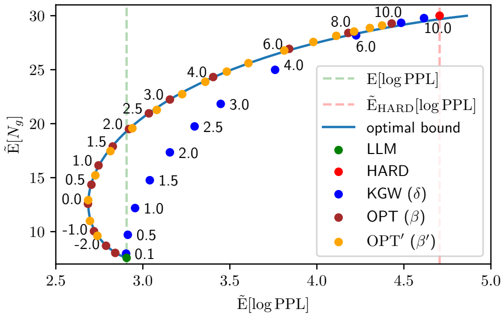

Figure 10. same as Figure 1, but now including the OPT′ watermark. Note that OPT and OPT′ hardly differ in terms of test-text trade-off, when text quality is defined in terms of expected log-perplexity. Error bars (vertical and horizontal) are never larger than the marker sizes.

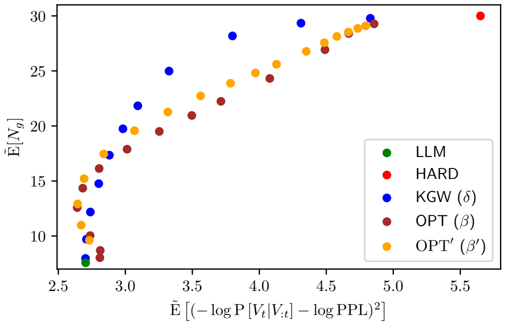

Figure 11. Test quality, measured as the expected number of green-list tokens, versus between-token variance of the log-perplexity, is shown for different watermarks. For completeness, the original language model without watermark is included (LLM). Error bars (vertical and horizontal) are never larger than the marker sizes. Note that, as expected, OPT' outperforms OPT, albeit marginally. Also note that the KGW watermark outperforms the optimized watermarks. This is not in contradiction with our method, as we did not optimize for this trade-off.

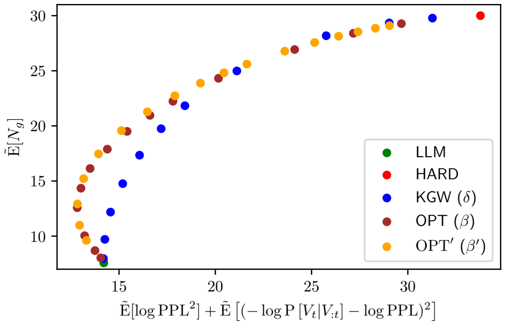

Figure 12. Test quality, measured as the expected number of green-list tokens, versus expected log-perplexity squared plus between-token variance of the log-perplexity, is shown for different watermarks. For completeness, the original language model without watermark is included (LLM). Error bars (vertical and horizontal) are never larger than the marker sizes. Note that, as expected, OPT' outperforms both OPT and KGW, as this trade-off is the optimization objective for which OPT' is optimized.