

How Good is Post-Hoc Watermarking With Language Model Rephrasing?

Pierre Fernandez^{*,1}, Tom Sander^{*,1}, Hady Elsahar¹, Hongyan Chang¹, Tomáš Souček¹, Valeriu Lacatusu¹, Tuan Tran¹, Sylvestre-Alvise Rebuffi¹, Alexandre Mourachko¹

¹FAIR, Meta Superintelligence Labs

*Core & Equal contributors.

Generation-time text watermarking embeds statistical signals into text for traceability of AI-generated content. We explore *post-hoc* watermarking where an LLM rewrites existing text while applying generation-time watermarking, to protect copyrighted documents, or detect their use in training or RAG via watermark radioactivity. Unlike generation-time approaches, which are constrained by how LLMs are served, this setting offers additional degrees of freedom for both generation and detection. We investigate how allocating compute (through larger rephrasing models, beam search, multi-candidate generation, or entropy filtering at detection) affects the quality-detectability trade-off. Our strategies achieve strong detectability and semantic fidelity on open-ended text such as books. Among our findings, the simple Gumbel-max scheme surprisingly outperforms more recent alternatives under nucleus sampling, and most methods benefit significantly from beam search. However, most approaches struggle when watermarking verifiable text such as code, where we counterintuitively find that smaller models outperform larger ones. This study reveals both the potential and limitations of post-hoc watermarking, laying groundwork for practical applications and future research.

Correspondence: tomsander@meta.com, pfz@meta.com

Code: <https://github.com/facebookresearch/textseal>



1 Introduction

Post-hoc text watermarking inserts an algorithmically-detectable signal into an existing text while remaining imperceptible to readers. It serves several goals, such as copyright protection or traitor tracing. Early approaches relied on hand-crafted modifications: synonym substitutions (Topkara et al., 2006c; Shirali-Shahreza and Shirali-Shahreza, 2008), grammatical transformations (Topkara et al., 2006b,a) or morphosyntactic alterations (Meral et al., 2009). Recent methods based on deep neural networks follow a post-hoc paradigm where the input is the original text and the output is the watermarked text, using one watermark embedder and one watermark extractor (Abdelnabi and Fritz, 2021), similarly to state-of-the-art post-hoc watermarking approaches for images, audio, and video. However, these methods are not effective, suffering from low capacity or unreliable detection, because they are highly constrained or easily broken by reversing the edits or synonyms.

With the emergence of large language models (LLMs), their popularization through ChatGPT (OpenAI, 2022), and growing concerns about potential risks (Crothers et al., 2022; Weidinger et al., 2022), generation-time text watermarking algorithms have been introduced to help detect AI-generated text (Aaronson and Kirchner, 2023; Kirchenbauer et al., 2023a). These algorithms have been deployed at scale, for instance, in Google’s Gemini with SynthID (Dathathri et al., 2024). Most LLM watermarks alter the next token selection, for example, by promoting a specific set of tokens depending on previous tokens and a secret key. Detection of the watermark in a text is then performed through a theoretically grounded statistical test that provides rigorous guarantees over the false positive rates. In addition, these methods achieve high watermark power while adding minimal latency by leveraging the entropy of the LLM to embed the signal. However, in contrast with previous approaches, these methods require control at generation time and therefore cannot be applied to existing text.

A natural idea to extend these methods to post-hoc watermarking is to employ a paraphrasing LLM to re-generate the text, allowing the watermark to be injected at inference time without further model training. This approach has been explored in data protection literature (Jovanović et al., 2025; Sander et al., 2025; Rastogi et al., 2025; Zhang et al., 2025; Lau et al., 2024). For example, “watermark

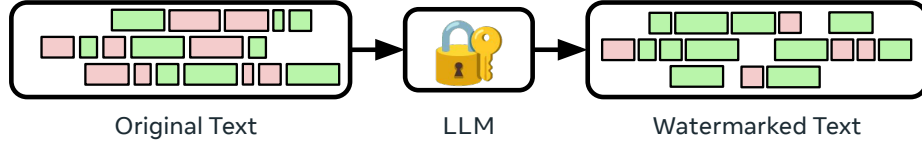


Figure 1 Post-hoc text watermarking through watermarked LLM rephrasing. We do empirical evaluations and analyze detection power, semantic fidelity, and correctness according to different design choices such as watermark scheme and available compute (through the paraphrasing model and the decoding strategy).

radioactivity” exploits the fact that watermark signals leave detectable traces when watermarked text is used by another model, enabling active training and context membership inference (Sander et al., 2025; Jovanović et al., 2025). However, to our knowledge, there is no thorough evaluation of how post-hoc watermarking through LLM paraphrasing performs across different data domains, such as prose versus verifiable text like code. Moreover, unlike generation-time watermarking, where watermark embedding should not delay the generation, post-hoc watermarking allows trading additional computation for a better quality-detectability trade-off. This opens up several axes to explore: model size, watermarking method, decoding strategy (and even generating multiple candidates), as well as detection strategy.

To address this gap, in this paper, we provide a comprehensive evaluation of post-hoc watermarking. We find that current state-of-the-art LLM watermarking schemes applied in a post-hoc setting perform well on open-ended text, such as Wikipedia articles and books, achieving high detection power while preserving fidelity. However, these methods are less effective on verifiable text like code, where the requirement to preserve correctness severely constrains paraphrasing freedom. Regarding design choices, we find that larger models better preserve semantics, while smaller models are necessary to effectively hide strong watermarks. Furthermore, the simplest Gumbel-max approach (Aaronson and Kirchner, 2023) dominates the Pareto frontier when classical random sampling is used. For other watermarking methods, we find that beam search decoding significantly improves the quality-detectability trade-off.

In short, our main contributions are:

- *A comprehensive evaluation:* We conduct the first large-scale study of post-hoc watermarking through LLM rephrasing (Figure 1) across diverse domains, demonstrating that while current methods are effective for open-ended text, they struggle with verifiable formats like code.
- *An analysis of design strategies:* We isolate the impact of model size, decoding, and watermarking schemes. For instance, we identify Gumbel-max as the robust Pareto-optimal choice (Figure 2).
- *An open-source research framework:* We release an easily modifiable codebase to facilitate research for post-hoc watermarking techniques.

2 Related work

2.1 Post-Hoc Text Watermarking

Early text watermarking altered text characteristics like characters or spacing (Brassil et al., 1995). Other methods modify grammatical or syntactical structures via pre-established rules (Topkara et al., 2005), including synonym substitution (Topkara et al., 2006c) and word reordering through passivization or topicalization (Topkara et al., 2006b,a; Meral et al., 2009). Text steganography follows similar principles (Winstein, 1998; Chapman et al., 2001; Bolshakov, 2004; Shirali-Shahreza and Shirali-Shahreza, 2008; Chang and Clark, 2014; Xiang et al., 2017). These edit-based systems typically exhibit low robustness and payload, e.g., 1-2 bits per sentence (Wilson and Ker, 2016). Deep learning methods have since been used for this task, for instance, with masked language models for steganography (Ueoka et al., 2021), infilling models (Yoo et al., 2023a), neural lexical substitution (Qiang et al., 2023), or encoder-decoder architectures (Abdelnabi and Fritz, 2021; Zhang et al., 2024; Xu et al., 2024).

2.2 Generation-Time Watermarking With Large Language Models

The first watermarks for machine-generated text date back to a method presumably used in Google Translate to filter automated translations from future training data (Venugopal et al., 2011). For LLM-generated text, two concurrent approaches appeared shortly after the release of ChatGPT. Kirchenbauer et al. (2023a) bias a subset of the vocabulary, while Aaronson and Kirchner (2023) alter the sampling via the Gumbel trick. Both use pseudorandom seeds generated from a secret key and preceding tokens, enabling lightweight detection through statistical tests without access to the model.

Subsequent work explores improved tests and multi-bit watermarking (Fernandez et al., 2023; Yoo et al., 2023b, 2024; Qu et al., 2024), position-dependent seeds (Christ et al., 2023; Kuditipudi et al., 2023), low-entropy optimizations (Lee et al., 2023; Christ et al., 2023; Huang et al., 2023), and semantic watermarks for better robustness (Liu et al., 2023; Liu and Bu, 2024; Fu et al., 2024; Hou et al., 2023, 2024). DiPMark (Wu et al., 2023) provides distortion-free Green-Red watermarks, and MorphMark (Wang et al., 2025) adaptively adjusts watermark strength based on the green token probability mass. WaterMax (Giboulot and Furon, 2024) generates several chunks of tokens from the original LLM distribution and selects outputs with high watermark scores, which ensures that the original LLM distribution is preserved. SynthID-Text (Dathathri et al., 2024) deploys tournament-based sampling in Google Gemini. Toolkits have also been introduced to benchmark these methods (Piet et al., 2023; Pan et al., 2024).

We provide a comprehensive description of the schemes evaluated in this work in subsection A.2.

2.3 Post-Hoc LLM Watermarks for Data Protection

Recent works apply LLM watermarks to training or evaluation data via paraphrasing, similar to what we study in this work. They exploit watermark radioactivity (Sander et al., 2024), i.e., the detectable traces left when watermarked text is used for training. Applications include detection of texts used in retrieval-augmented generation (RAG) (Jovanović et al., 2025), benchmark contamination detection (Sander et al., 2025), and training data copyright (Zhang et al., 2025). Only Waterfall (Lau et al., 2024) evaluates post-hoc watermarking through LLM paraphrasing, focusing on code (MBPP) and natural text (C4 and arXiv) for provenance detection.

These works demonstrate the utility of such methods for data protection, but do not evaluate it as a general watermarking method, nor do they characterize failure modes across text types and settings.

3 Method

Post-hoc watermarking via paraphrasing. The key idea is to paraphrase the input text using an LLM while applying watermarking during generation, as depicted in Figure 1. The pipeline is as follows:

1. Split input text into chunks (sentences or paragraphs).
2. For each chunk, prompt an LLM to produce a paraphrase, given specific instructions (e.g., preserve named entities, minimal lexical change) and some previous context.
3. During decoding, sample the next token with watermarking: for instance, favor tokens in a “greenlist” according to a strength parameter or any of the methods described in subsection 2.2.
4. Aggregate watermark scores across chunks and compute the detector statistic.

3.1 LLM Watermarking

Next token generation. At each decoding step, the LLM computes logits $\ell \in \mathbb{R}^{|\mathcal{V}|}$ over the vocabulary \mathcal{V} conditioned on the preceding context. These logits are converted into a probability distribution $\mathbf{p} = \text{softmax}(\ell/T)$ via temperature scaling by T .

Watermarked sampling. Watermarking modifies the decoding process rather than sampling directly from \mathbf{p} . This is done depending on a window \mathbf{w} of the k preceding tokens and a secret key s . A pseudorandom function $\text{PRF}(\cdot)$ combines \mathbf{w} , s , and a candidate token v to produce a value used for biasing the sampling of token v . For instance, in the green-red method (Kirchenbauer et al., 2023a), we

compute $\text{PRF}(\mathbf{w}, s, v)$ for all tokens $v \in \mathcal{V}$ to partition the vocabulary into a “greenlist” and a “redlist,” then bias the logits by adding δ to green tokens. Other schemes alter the sampling mechanism directly (e.g., with the Gumbel-max trick (Aaronson and Kirchner, 2023)). This embeds a statistical signal into the generated text that is imperceptible to readers but detectable by the detector. We provide a comprehensive description of the watermarking schemes evaluated in this work in [subsection A.2](#).

Detection and statistical test. Detection recomputes the pseudorandom function for each group of tokens and aggregates the results into a test statistic. Under the null hypothesis \mathcal{H}_0 “not watermarked with this specific scheme”, the statistic should follow a known distribution, yielding a p -value for the probability of observing such a score or a more extreme score by chance. Under the alternative hypothesis \mathcal{H}_1 “watermarked text with this specific scheme”, the statistic should deviate significantly enough and lead to low p -values. We flag text as watermarked if the p -value is below a detection threshold α , which controls the false positive rate (FPR).

For instance, if K (the test statistic here) is the observed number of green tokens out of N scored tokens and this statistic follows under \mathcal{H}_0 a Binomial distribution with parameters N and γ (expected green ratio, typically 0.5), then $p\text{-value} = \mathbb{P}(X \geq K \mid X \sim \text{Binomial}(N, \gamma))$. As an example, if we observe $K = 65$ green tokens out of $N = 100$ with $\gamma = 0.5$, the p -value is approximately 10^{-3} ; to flag this text as watermarked, we would need a FPR of $\alpha \geq 10^{-3}$.

An important practical consideration is *deduplication*: because the pseudorandom function depends on the preceding k tokens, repeated n -grams generate identical hashes. This violates the statistical independence of the scores used when assuming the distribution of the test statistic under \mathcal{H}_0 . Aggregating scores only over unique watermark windows within the text (Kirchenbauer et al., 2023a; Fernandez et al., 2023) is a good way to mitigate this issue and to ensure valid p -values. But deduplication alone may not guarantee valid statistical tests because natural language inherently favors certain n -grams over others. This bias can cause p -values to be artificially low or high even for unwatermarked text. To address this, we ensure that under \mathcal{H}_0 , p -values are approximately uniform on $U(0, 1)$. In practice, we test many candidate secret keys and select one for which the empirical p -value distribution on unwatermarked text is close to uniform. We provide more details in [subsection A.1](#).

3.2 Compute-Driven Flexibility

Operating in a post-hoc setting offers more flexibility than in-generation watermarks, which must comply with live-serving constraints such as latency, memory, and decoding speed (Dathathri et al., 2024). Here, we can use large or small models, run higher temperatures or beam search instead of simple sampling, and even employ multi-candidate selection like WaterMax (Giboulot and Furon, 2024), as a function of available compute and quality targets. We detail the specific higher-compute decoding (and detecting) methods we test below.

Beam search for watermarking. We use a beam search approach to improve the quality-detectability trade-off. Specifically, we maintain B candidate sequences and expand each beam with V candidates sampled from the watermarked probability distribution $\mathbf{p}_{\text{wm}}(\cdot \mid x_{<t})$, which is derived by applying the watermarking scheme (e.g., Maryland, SynthID, DIPMark, MorphMark) to the base model’s logits. To select the top- B beams at each step, we score candidates using log-probability under a reference distribution. We explore two scoring variants: *unbiased scoring* uses the original model probabilities $\mathbf{p}_{\text{orig}}(\cdot \mid x_{<t})$ to favor sequences with lower perplexity relative to the base model, prioritizing text quality; *biased scoring* uses the watermarked probabilities $\mathbf{p}_{\text{wm}}(\cdot \mid x_{<t})$ to favor sequences that are most probable under the watermarked distribution, potentially yielding stronger watermark signals. We note that the latter was also explored for generation-time watermarking in (Kirchenbauer et al., 2023b). For candidate generation, we consider both deterministic beam search (selecting top- V tokens by \mathbf{p}_{wm}) and stochastic beam search (sampling V tokens from \mathbf{p}_{wm}).

Multi-candidate selection with WaterMax (Giboulot and Furon, 2024). Rephrasing is done chunk by chunk, with chunks of L tokens. For each chunk, we generate m candidates $\{\tilde{y}^{(1)}, \dots, \tilde{y}^{(m)}\}$ without applying any logit bias. We then select the candidate that naturally maximizes the watermark score: $y^* = \arg \max_y \text{Score}_{\text{wm}}(y)$. This method is “distortion-free” regarding the sampling distribution but is typically infeasible in standard API usage due to cost (generating L times more tokens per chunk).

Entropy-aware detection. At detection time, we compute the entropy H_t of the model’s predicted distribution at each token position t in the watermarked text $H_t = -\sum_{v \in \mathcal{V}} p_t(v) \log p_t(v)$ where $p_t(v)$ is the model probability for token v given the prefix up to $t-1$ in the watermarked text. We then apply an entropy filter: a token is included in the watermark score only if H_t exceeds a chosen threshold τ , similar to what is done in Lee et al. (2023) for code. This focuses detection on high-entropy positions, where watermarking is more effective, and ignores low-entropy tokens. All entropy computations are performed on the watermarked text, as the original text is not available at detection time. As a result, these entropy values differ from those at generation-time, when the model was conditioned.

4 Experiments

We conduct a comprehensive evaluation of post-hoc watermarking to assess: (1) *detection power* (the ability to reliably identify watermarked text, measured through p -values) and (2) *fidelity* (the preservation of meaning and quality through paraphrasing). In subsection 4.1, we detail the experimental setup, including datasets, models, watermarking schemes, decoding strategies, and evaluation metrics. In subsection 4.2, we compare watermarking methods on the quality-detectability Pareto frontier, finding that Gumbel-max dominates under standard sampling. In subsection 4.3, we show that larger models preserve semantics better, but small/mid-size models offer sweet spots, with stronger watermark signals due to higher entropy. In subsection 4.4, we demonstrate that beam search improves the Pareto frontier for all applicable methods. In subsection 4.5, we evaluate entropy-aware detection and find only modest gains. In subsection 4.6, we investigate post-hoc watermarking on code, revealing that correctness constraints limit detectability. In subsection 4.7, we evaluate cross-lingual robustness and, in subsection 4.8, the impact of chunking on long documents.

4.1 Experimental Setup

Datasets. We evaluate on three diverse corpora to capture different text characteristics and use cases:

- *Books*: Passages in English from the Gutenberg dataset (Project Gutenberg, 2025), containing books. These represent longer-form, literary texts with complex sentence structures.
- *Wikipedia*: The beginning of randomly sampled Wikipedia articles. These represent factual, encyclopedic texts with a formal writing style and different languages.
- *Code*: Programming problems from the HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) benchmarks, consisting of Python functions with accompanying unit tests.

Paraphrasing models and prompts. We experiment with several dense open-weight instruct LMs as paraphraser, spanning different families and sizes to study the impact of model scale on paraphrase quality and watermark preservation. Specifically, we use the instruct version of **LLaMA-3** (Dubey et al., 2024) (1B, 3B, 8B, 70B), **Gemma-3** (DeepMind, 2025) (1B, 4B, 12B, 27B), **Qwen-2.5** (Team, 2024) (0.5B, 1.5B, 3B, 7B, 14B, 32B), and **SmolLM-2** (Allal et al., 2025) (135M, 360M, 1.7B).

To ensure the model outputs the watermarked text directly and without any comments, we use a structured system prompt and prefill the assistant’s response with a prefix (e.g., “Rephrased text:”). An example prompt used is:

“You are a text rephrasing assistant. You must rephrase the given text while strictly preserving its original meaning, style, and structure. You must output only the rephrased text, with no explanations or commentary.”

Chunking strategy. Large documents may be difficult to rephrase in a single pass, or may even exceed the model’s context window, so we split texts into chunks. We compare two modes: *full-context*, which processes the entire document at once, and *context-aware chunking*, which processes each chunk separately while prepending one or two previously rephrased chunks as context to preserve coherence.

Watermarking schemes. We evaluate multiple watermarking methods (detailed in subsection A.2), all with window size 2, varying their hyperparameters to explore their detection-quality trade-offs.

- *Green-Red List* (Kirchenbauer et al., 2023b): We vary the bias $\delta \in \{1.0, 2.0, 4.0\}$.

- *DiPMark* (Wu et al., 2023): We vary the reweighting parameter $\alpha \in \{0.2, 0.3, 0.4\}$.
- *MorphMark* (Wang et al., 2025): We fix $\kappa = 10$ and vary $p_0 \in \{0, 0.05, 0.1, 0.2\}$.
- *SynthID-Text* (Dathathri et al., 2024): We vary the number of tournaments $k \in \{10, 20, 30\}$. We also compute the p -values from weighted and non-weighted scores as described in subsection A.2.
- *Gumbel-max* (Aaronson and Kirchner, 2023): Fixed by temperature and top- p .

Choice of private key. We address the private key issue described in section 3 by testing 50 candidate keys. For each combination of tokenizer, watermarking method (and depth for SynthID), we evaluate the keys on non-watermarked texts to estimate the p -value distribution under H_0 . We then select the key that minimizes the deviation from $U(0, 1)$, as measured by the Kolmogorov–Smirnov statistic. Next, we verify that the theoretical FPRs for these keys match the empirical FPRs on 1.5M Wikipedia documents (they contain up to 100k tokens, with an average of 715 tokens). This ensures that the subsequent results under H_1 reflect the intrinsic performance of each scheme rather than artifacts of a particular key choice. Further details are provided in Table 6.

Entropy aware detection. For all methods, we evaluate entropy-aware detection by computing per-token entropy with the same model as used for rephrasing. At detection, we only score tokens with entropy $\geq \tau$, with $\tau \in \{0, 0.2, 0.4, 0.6, \dots, 1.8, 2.0\}$. We note that we only pass the watermarked text to compute the per-token entropy without the original text as context.

Decoding strategies. We explore different decoding approaches:

- *Nucleus sampling* (Holtzman et al., 2019). We vary temperature $T \in \{0.7, 1.0, 1.2\}$ and top- $p \in \{0.9, 0.95, 0.99\}$. Higher temperature or top- p should increase entropy at the cost of faithfulness.
- *Beam search* (Sutskever et al., 2014), as detailed in section 3. We test beam widths $B \in \{3, 5, 10\}$ with the same number of candidates per beam. We experiment with selection criteria based on either the non-watermarked likelihood or the watermarked model’s likelihood, and use stochastic (temperature 1 and top- p 0.95) or deterministic selection within beams.
- *WaterMax* (Giboulot and Furon, 2024), as detailed in section 3. We generate $L \in \{4, 8, 16\}$ non-watermarked candidate tokens per draft $m \in \{4, 8\}$ and select the one with the highest number of green tokens, and repeat.

Evaluation metrics. We employ the following metrics to evaluate performance:

- *Detection*: We compute the p -value of the corresponding detection test, which can be read as the probability that a watermark score at least as high would happen if the text were not watermarked.
- *Quality*: We evaluate rephrasing quality along three axes. *Semantic similarity*: We use the BERTScore (Zhang et al., 2020) to measure meaning preservation between original and rephrased texts. *Naturalness*: We compute the average cross entropy of the rephrased text using Mistral-7B-v0.3 (Jiang et al., 2023), conditioned on the prefix “*This is a rephrased version of [ORIGINAL TEXT]:*”, capturing if the output is fluent and plausible. *Length ratio*: We report the character-level length ratio between rephrased and original text, which should remain close to 1.
- *Functional Correctness (Code)*: For code datasets, we measure the percentage of paraphrased functions that pass their unit tests (Pass@1).

In Sections 4.2, 4.3, 4.5, 4.4 and 4.8, each reported value corresponds to the median over 100 passages of approximately 800 characters (~ 200 tokens) sampled from Charles Dickens’ novels in the Project Gutenberg dataset (Project Gutenberg, 2025). We keep only rephrasings where the output length remains comparable to the original (length ratio $\in [0.75, 1.25]$), and report results only for settings where at least 70% of passages satisfy this constraint.

As detailed in Appendix C, output length distributions are remarkably consistent across watermarking schemes and are primarily determined by the base model rather than the watermarking method. While output length is arguably a quality signal in itself, we invite the reader to examine these distributions in the appendix, allowing the main analysis to focus on semantic fidelity and detection power within valid generation bounds.

Table 1 Qualitative comparison on literary text. A sample from *A Christmas Carol* by Charles Dickens, rephrased using Llama-3.2-3B-Instruct with the Gumbel-max scheme, $T = 0.7$ and $\text{top-p} = 0.95$. The output maintains the original tone and structure while achieving high semantic similarity and good watermark detectability.

Original Text	Watermarked Rephrasing
“You don’t believe in me,” observed the Ghost.	“You’re not convinced of my existence,” said the Ghost.
“I don’t,” said Scrooge.	“I am not,” replied Ebenezer Scrooge.
“What evidence would you have of my reality beyond that of your senses?”	“What evidence do you have to confirm my reality beyond the limits of your perceptions?”
“I don’t know,” said Scrooge.	“I don’t know,” Scrooge replied.
“Why do you doubt your senses?”	“Why do you question the reliability of your senses?”
“Because,” said Scrooge, “a little thing affects them. A slight disorder of the stomach makes them cheats. You may be an undigested bit of beef, a blot of mustard, a crumb of cheese, a fragment of an underdone potato. There’s more of gravy than of grave about you, whatever you are!”	“Because,” Scrooge said, “a minor aberration can significantly impact them. A slight irregularity in digestion can cause them to be misled. You may be an undigested morsel of meat, a splash of mustard, a fragment of a partially cooked potato. There’s more of gristle than of flesh about you, regardless of what you are!”
Scrooge was not much in the habit of cracking jokes, nor did he feel, in his heart, by any means waggish then. The truth is, that he tried to be smart, as a means of distracting his own attention, and keeping down his terror; for the spectre’s voice disturbed the very marrow in his bones. To sit, staring at those fixed glazed eyes, in silence for a moment, would play, Scrooge felt, the very deuce with him.	Scrooge was not accustomed to indulging in humor, nor did he feel that he was particularly witty at any point. The truth was, that he tried to appear intelligent, as a means of diverting his own attention and suppressing his fear; as a way to distract himself from the unsettling presence of the spectre. To sit, staring at those fixed, unblinking eyes, in silence for a moment, would play, Scrooge felt, a cruel and unforgiving reminder of the truth.
Stats: SBERT: 0.904 • Ratio: 1.07 • Perplexity: 2.61 • $p\text{-val: } 1.7 \times 10^{-5}$	

4.2 Quality-Detection Trade-Off

We first evaluate the fundamental trade-off between rephrasing quality and detection strength in Figure 2, using Llama-3.2-3B-Instruct with random sampling. Each point corresponds to a distinct parameter configuration as described in subsection 4.1. Optimal methods occupy the top-right region of the plot, combining high text quality with strong detectability. The Gumbel-max method appears to dominate this Pareto frontier, followed by MorphMark and SynthID.

We also present a qualitative example in Table 1, where we rephrase text using Llama-3.2-3B-Instruct with the Gumbel-max watermark with $T = 0.7$ and $p = 0.95$. The rephrasing maintains the original tone, structure, and factual content. Additional examples are provided in Table 8 of Appendix B.

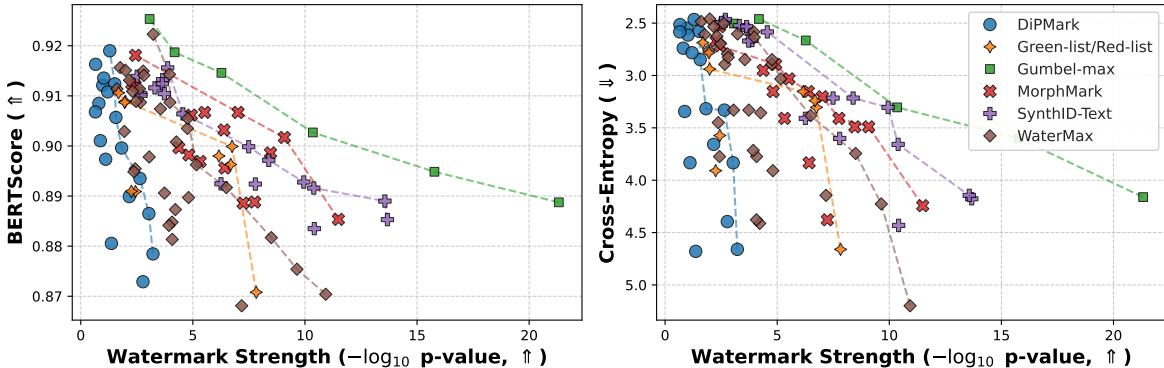


Figure 2 Pareto fronts of watermarking methods showing the trade-off between quality and watermark strength using Llama-3.2-3B-Instruct. Each point corresponds to a different parameter configuration, with values representing medians across 100 rephrased passages. Experimental details are given in Sections 4.1 and 4.2.

4.3 Impact of Model Family and Size

We investigate whether larger models are better suited for post-hoc watermarking. Figure 3 shows the trade-off curves by model family (SmolLM, Gemma, Llama, Qwen) and size. Regarding performance, we observe that larger models preserve semantics better: across all families, increasing model size (darker points) shifts the clusters upward. For example, within the Llama family, Llama-3.1-8B (dark green) attains lower cross entropy than the 3B and 1B variants (lighter teal) at comparable watermark strengths. However, larger models struggle to reach low p -values compared to smaller models: they populate the left rather than the right of the plots. Gemma 3 models are essentially absent from the high-detectability region, as their outputs naturally exhibit very low entropy, even for the smallest sizes. Overall, when strong watermarking is required, only small models appear on the frontier, whereas larger models dominate in the lower-strength regime. We see in subsection 4.6 that the issue with large models persists even if the temperature is increased.

Note that these results are conditioned on the length filtering described in subsection 4.1; thus, the presence of a data point implies enough successful valid generation for the corresponding combination of model/scheme/parameters. For instance, smaller models with stronger watermarks fail to meet length constraints more frequently. We refer the reader to Appendix C to assess the generation stability of each configuration, which cannot be inferred directly from the figure.

4.4 Decoding Strategies: Beam Search vs. Sampling vs. WaterMax

We explore in Figure 4 whether beam search can find better watermarked sequences than random sampling. We watermark with Llama-3.2-3B-Instruct, and decode with or without beam search, for each suitable watermarking method: Green-red, Synth-ID, MorphMark, and DiPMark (since Gumbel-max is deterministic, beam search cannot be applied with it).

We observe that beam search consistently improves the Pareto frontier, *especially with biased scoring* (see section 3): it shifts the results upward, substantially improving rephrasing quality at a fixed watermark strength. Notably, while Gumbel-max sampling appears to be the best option when used with random sampling in Figure 2, other methods can be substantially improved through beam search.

WaterMax (Giboulot and Furon, 2024) results are shown in brown in Figure 2. Surprisingly, we found that WaterMax achieved weak detectability across the hyperparameters we tested, making it difficult to achieve strong watermarking detection in the post-hoc setup.

4.5 Entropy-Aware Detection

Instead of scoring every token, we now filter out low-entropy tokens at detection time, as motivated in section 3 and detailed in subsection 4.1. In Figure 5, we fix the rephrasing model to Llama-3.2-3B and, for each method and watermarking configuration, test whether there exists an entropy threshold

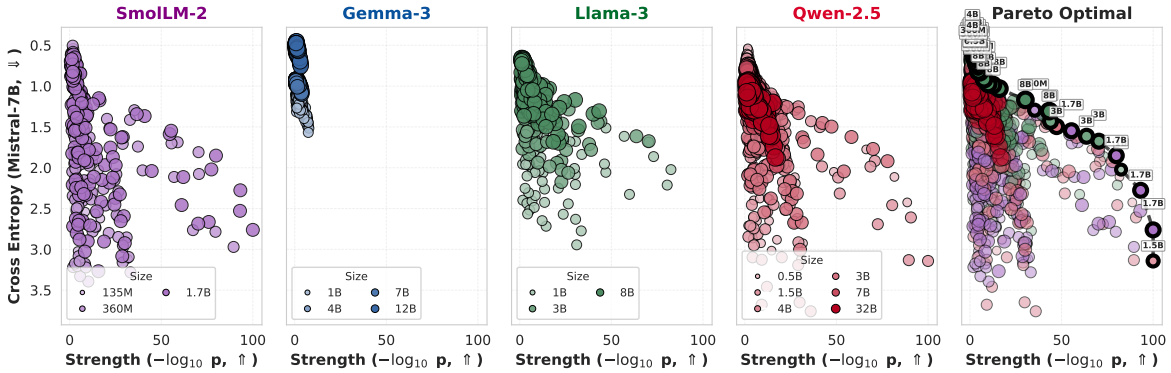


Figure 3 Impact of model family and size. Cross Entropy vs. watermark strength. Larger models improve quality for a given watermark strength, but small models are necessary to reach high strengths. All families are comparable, except Gemma-3 which is not suitable. Experimental details are given in Sections 4.1 and 4.3.

that improves detectability by more than 5% on at least 50 of the 100 texts of Charles Dickens.

The left plot reports, for each watermarking method, the proportion of configurations for which such a threshold exists. We see that most methods have only 30–40 successful configurations. The middle plot shows, for those successful configurations, the actual gain in detection performance, which never exceeds 20%. The right panel presents the corresponding non-aggregated statistics.

We see that WaterMax is different: any entropy threshold degrades detectability. This is because WaterMax optimizes the watermark score at the *sentence* level rather than the token level; filtering by token entropy does not expose additional signal, it only reduces the number of scored tokens and thus the available statistical evidence. One other and related property of WaterMax is that it cannot be used for active dataset inference through radioactivity. This is explained in detail in [Appendix C](#).

4.6 Post-hoc Code Watermarking

Experimental details. We now conduct experiments on HumanEval and MBPP to evaluate the feasibility of post-hoc watermarking on structured and verifiable text, where watermarking requires preserving the syntax and the function of the code.

Data preparation and preprocessing. For each problem instance, we concatenate the provided prompt

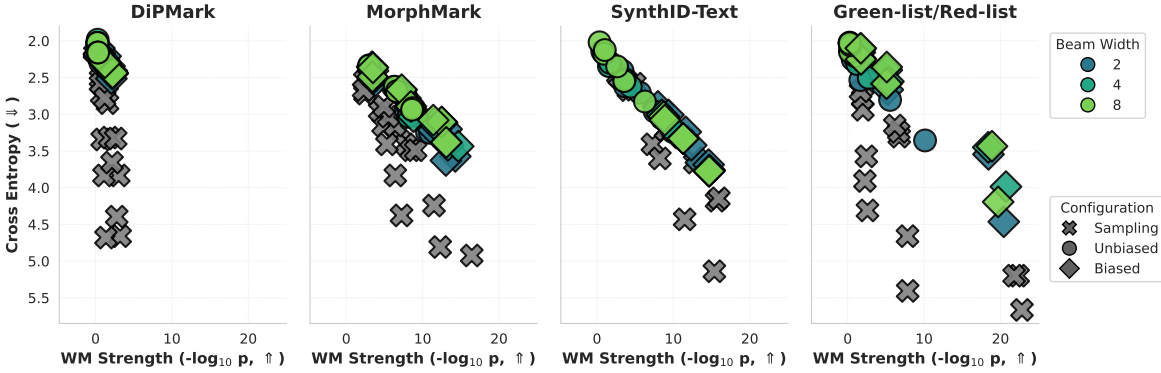


Figure 4 Beam search improves the Pareto frontier. Cross entropy vs. watermark strength for suitable methods. Beam search, especially with biased scoring (see [section 3](#)), shifts the frontier upward, substantially improving rephrasing quality at a fixed watermark strength. Experimental details are given in Sections [4.1](#) and [4.4](#).

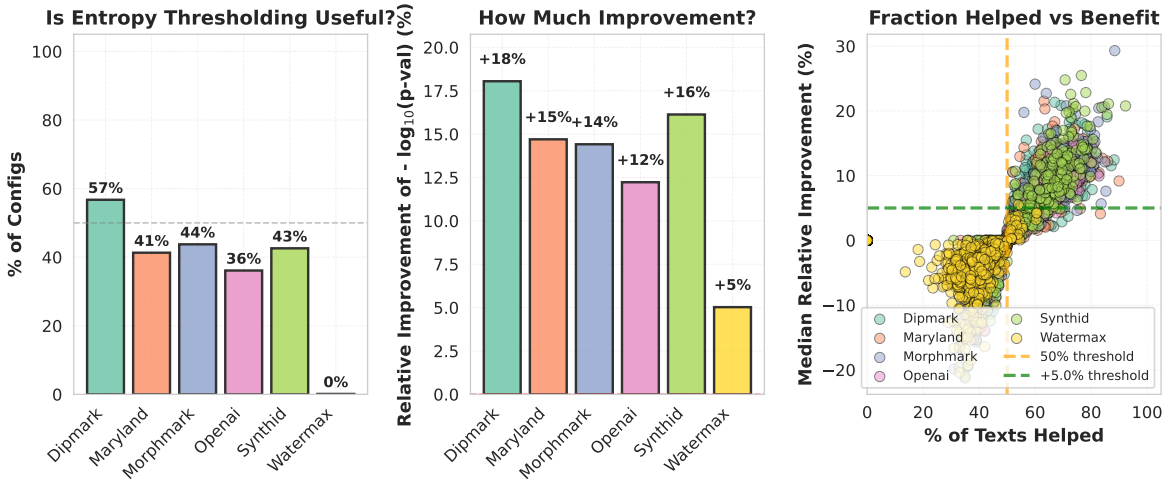


Figure 5 Effect of entropy-aware detection. *Left:* Share of configurations for which some threshold improves detection by more than 5% on at least half of the texts. *Middle:* Median relative improvement for those configurations. *Right:* For every configuration, fraction of texts helped at optimal threshold vs. median improvement; dashed lines mark the 50% and +5% criteria. Experimental details are given in Sections [4.1](#) and [4.5](#).

Table 2 Qualitative example of code watermarking on a sample from MBPP (using Llama-3.1-8B at temperature 1.4 with Gumbel-max watermarking, top- $p = 0.95$). The post-hoc rephrasing here refactors variable names to ensure detection (low p-value) while maintaining correctness.

Original Text	Watermarked Rephrasing
<pre># A python function to identify non-prime numbers. import math def is_not_prime(n): result = False for i in range(2,int(math.sqrt(n)) + 1): if n % i == 0: result = True return result</pre>	<pre># A python function to identify non-prime numbers import math def is_not_prime(number): verdict = False for divisor in range(2, int(math.sqrt(number)) + 1): if number % divisor == 0: verdict = True return verdict</pre>
p_{value} : 0.12 • Tokens: 48 • Correct: ✓	p_{value} : 3.29×10^{-3} • Tokens: 57 • Correct: ✓

(e.g., function signature and docstring) and the canonical solution. We make sure that the last top-level function definition of the code is the target for execution by moving it if necessary. We show in App. B.1 an example of a HumanEval task. Overall, this dataset comprises 164 HumanEval and 974 MBPP code samples, with average lengths of 180 and 80 tokens, respectively. The code samples range from approximately 20 to 500 tokens in length.

Watermarking pipeline. The watermarking process is applied only to the code (we exclude unit tests to preserve the evaluation). Since watermarking may inadvertently rename function identifiers, we implement a post-processing step that parses the output and restores the original entry point name required by the benchmark test harness. At test time, we concatenate the watermarked code with the test harness to evaluate functional correctness.

Metrics for code watermarking. First, we measure *functional correctness*: the fraction of watermarked code that passes the tests (pass@1). If pass@1 is low, one could regenerate with different random seeds until obtaining functional code, but this assumes tests are available and it is cumbersome; ideally, pass@1 should be high. Second, we measure the *detection power*: the true positive rate (TPR) at a fixed false positive rate (FPR), and/or the distribution of $\log_{10} p$ -values. These can be computed among all rephrased code, or more interestingly, among only the ones that passed the tests. In practice, the latter is a better indicator of the watermark power, since codes that do not pass tests often exhibit degenerative patterns (repeated code or text at the end) which artificially increase the TPR.

Comparing watermarking methods. We compare the watermarking methods using Llama-3.1-8B. Figure 6 shows the trade-off between functional correctness (pass@1) and detection power (TPR at

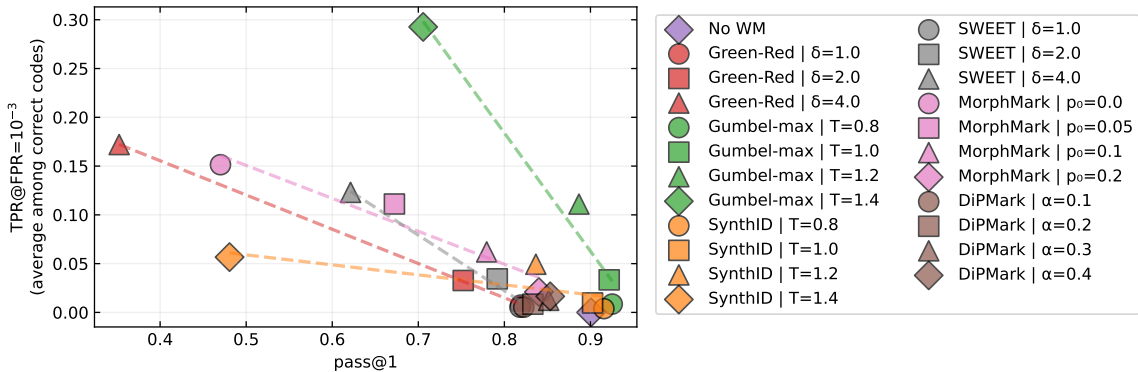


Figure 6 Comparison of methods on code using Llama-3.1-8B. We report pass@1 (functional correctness) vs. TPR at $\text{FPR}=10^{-3}$ among correct codes, averaged over HumanEval and MBPP. Different markers and colors correspond to different watermarking methods and their respective hyperparameters. Consistent with our findings in literary English, Gumbel-max watermarking achieves the best trade-off between utility and detectability.

Table 3 Post-hoc watermarking on code with Llama models using Gumbel-max (top- $p=0.95$). We report pass@1, median \log_{10} p-value (\uparrow), and TPR at FPR= 10^{-3} for different model sizes and temperatures (T). Metrics are shown over all samples (“All”) and restricted to samples where the watermarked code passes functional tests (“Passed”). Increasing model size increases pass@1 but decreases detection power, with smaller models achieving better Pareto optimality.

Model	$T = 1.0$						$T = 1.2$						$T = 1.4$					
	pass@1	$-\log_{10}$ p-val		TPR@ 10^{-3}			pass@1	$-\log_{10}$ p-val		TPR@ 10^{-3}			pass@1	$-\log_{10}$ p-val		TPR@ 10^{-3}		
		All	Passed	All	Passed			All	Passed	All	Passed			All	Passed	All	Passed	
Llama-3.2-1B-Instruct	0.59	0.60	0.51	0.03	0.01		0.44	1.32	0.88	0.22	0.07		0.16	9.10	1.56	0.72	0.29	
Llama-3.2-3B-Instruct	0.81	0.60	0.56	0.01	0.01		0.73	0.87	0.75	0.08	0.05		0.53	1.84	1.00	0.34	0.13	
Llama-3.1-8B-Instruct	0.92	0.74	0.74	0.04	0.03		0.89	1.16	1.11	0.12	0.11		0.71	2.31	1.83	0.41	0.29	
Llama-3.3-70B-Instruct	0.92	0.38	0.37	0.00	0.00		0.92	0.41	0.40	0.00	0.00		0.90	0.44	0.44	0.00	0.00	

FPR= 10^{-3} among correct codes) for each method. Gumbel-max watermarking achieves the best Pareto frontier, offering the most favorable trade-off between utility and detectability. SynthID performs well in moderate temperature regimes but breaks down at higher temperatures, which limits its ability to achieve high TPR values. MorphMark and DiPMark show competitive performance at intermediate operating points, while Green-Red and SWEET exhibit steeper degradation in pass@1 as watermark strength increases. Overall, no method achieves both high pass@1 and high TPR simultaneously, confirming the inherent difficulty of watermarking code while preserving functionality.

Comparing models. We compare model families and sizes using Gumbel-max watermarking (top- $p=0.95$). Figure 7a shows that larger models lead to lower detection power (lower $-\log_{10}$ p-values), as their outputs are less entropic and leave less room for watermarking. Table 3 confirms this: Llama-70B achieves median $-\log_{10}$ p-values around 0.5-0.7, while smaller models reach values near 2 at $T = 1.4$. However, higher detection power in smaller models comes at the cost of functional correctness; for example, at $T = 1.4$, Llama-8B achieves a TPR of 0.29 among correct codes but a pass@1 of only 0.71.

To further investigate this trade-off, we sweep the temperature in Figure 7b. Interestingly, the smaller 8B model achieves a better Pareto frontier than the larger 70B model, attaining higher detection power at equivalent pass@1 levels. This suggests that while larger models produce higher-quality code, their reduced entropy limits watermark effectiveness, making smaller models more suitable for post-hoc code watermarking when balancing functionality and detectability.

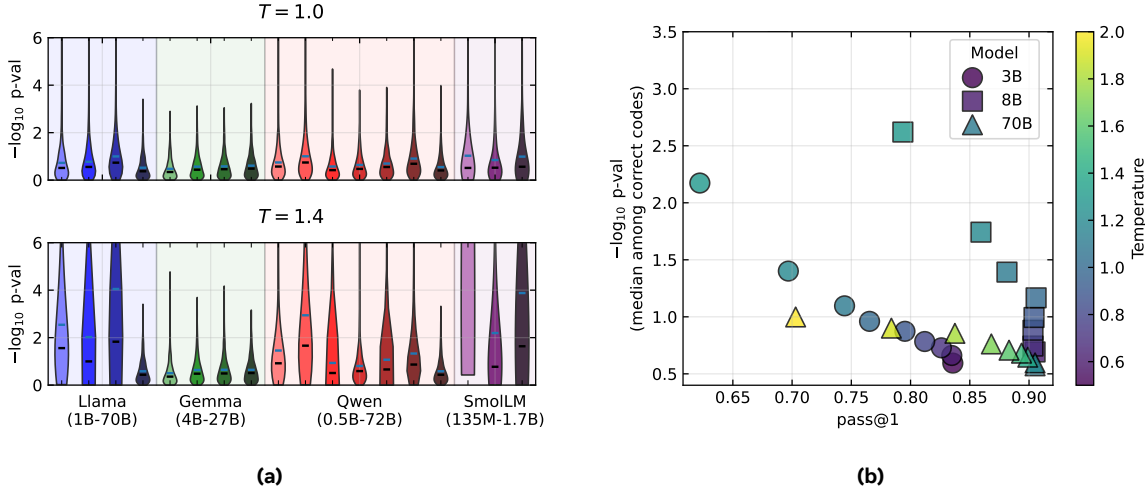


Figure 7 Evaluation of post-hoc watermarking across models using Gumbel-max (top- $p=0.95$). (a) Distribution of $-\log_{10}$ p-value for watermarked codes that pass the functional correctness test (bars indicate mean and median). Different colors represent model families, with darker shades indicating larger models. Corresponding values for Llama models are detailed in Table 3. (b) Pass@1 and detection power (median $-\log_{10}$ p-value among correct codes) for Llama 3B, 8B and 70B models across temperatures. Different model families and sizes behave differently in terms of detection power at fixed temperatures, and interestingly, smaller models can achieve better Pareto optimality than larger models.

Table 4 Evaluation of post-hoc watermarking for different languages. We report the median \log_{10} p-value and the SBERT score for semantic similarity, at varying temperatures of the generation. While not impossible, watermarking languages other than English comes at a steeper cost on semantic quality.

Temperature	en		es		fr		ru	
	$-\log_{10} p$	SBERT	$-\log_{10} p$	SBERT	$-\log_{10} p$	SBERT	$-\log_{10} p$	SBERT
0.80	1.832	0.960	1.608	0.949	0.810	0.946	1.355	0.908
1.00	3.813	0.954	4.408	0.940	2.809	0.937	4.976	0.888
1.20	8.147	0.946	20.098	0.921	19.705	0.911	23.689	0.836

4.7 Multi-Lingual Robustness on Wikipedia

To evaluate the generalization of post-hoc watermarking beyond English, we apply the method to paragraphs of Wikipedia articles in multiple languages. In the following, we use Llama-3.1-8B-Instruct, with Gumbel-max watermarking with varying temperature settings (0.8, 1.0, 1.2) and fixed top- $p = 0.95$. Table 4 compares the performance across English (en), Spanish (es), French (fr), and Russian (ru). To not conflate the effect of the number of tokens used for scoring with language, we only consider outputs that contain between 400 and 600 tokens, and aggregate results over 1k samples.

The results indicate a performance gap: while non-English languages (Spanish, French, Russian) can be watermarked, they suffer from a steeper trade-off. Achieving high detection strength in these languages requires a larger sacrifice in semantic quality compared to English. This suggests that the logits of the paraphrasing model are less easy to manipulate in languages for which the model was not primarily trained, since most training data is in English. It may also reflect the fact that our semantic similarity metric (SBERT) is more accurate for English than for other languages.

4.8 Impact of Chunking Strategy on Long Documents

To assess the impact of chunking as described in section 3, we compare full-context processing with context-aware chunking (500-token chunks, with up to 1000 context tokens from previously rephrased chunks) on documents of varying length (500-4000 tokens), using Dickens novel excerpts and Llama-3.2-3B-Instruct.

For each document length, results are averaged over 5 independent texts. Table 5 shows that full-context processing increasingly summarizes the content, while chunking better preserves length and yields stronger detection at better semantic similarity. This indicates that context-aware chunking is crucial for reliable watermarking of important document lengths.

5 Conclusion

We presented a comprehensive evaluation of post-hoc text watermarking via LLM rephrasing, a paradigm that enables embedding traceable signals into existing text. Unlike generation-time constraints, the post-hoc setting allows the allocation of additional compute to optimize the trade-off between text quality and watermark detectability.

Our experiments yielded several findings. First, the simplest Gumbel-max scheme (Aaronson and Kirchner, 2023) achieved better trade-offs than all other tested methods under random sampling. We note that in the generation-time watermarking literature, Gumbel-max is criticized for its deterministic

Size	Length Ratio		Detection ($-\log_{10} p$)		Similarity	
	Full	Chunked	Full	Chunked	Full	Chunked
500	0.91	0.91	19.3	19.3	0.94	0.94
1500	1.15	0.97	49.5	148.8	0.85	0.91
2500	0.78	0.98	56.8	195.5	0.91	0.91
4000	0.70	0.86	24.9	198.3	0.84	0.91

Table 5 Context-aware chunking vs. full-context. Comparison of full-context and context-aware chunked rephrasing (500-token chunks with up to 1000 context tokens) on documents of varying length, averaged over 5 excerpts per length. Experimental details are given in subsection 4.1 and subsection 4.8.

nature: it fixes the randomness of generation, so identical prompts produce identical outputs, which can be problematic in production. In our post-hoc setting, however, this limitation is less consequential. We also observe that all other schemes benefit substantially from beam search. Second, to achieve high watermark strength, smaller, more entropic models outperformed larger models run at high temperature. Third, entropy filtering at detection time provided only marginal gains while introducing additional complexity, making it a less practical option in typical deployment scenarios.

A limitation of our study, and of the field at large, is the reliance on automated metrics such as perplexity and BERTScore to approximate text quality and semantic preservation. Much like in image and audio watermarking, these proxies cannot fully capture the nuances of human perception, particularly across different languages where robust human evaluations remain essential. To address this limitation, we extended our analysis to post-hoc code watermarking, where execution-based correctness provides an objective, ground-truth measure of utility. Our results reveal that while watermarking is feasible in this domain, strict correctness constraints reduce the available capacity for watermark embedding compared to natural language.

References

- Scott Aaronson and Hendrik Kirchner. Watermarking GPT outputs, 2023.
- Sahar Abdelnabi and Mario Fritz. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 121–140. IEEE, 2021.
- Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Lewis Tunstall, Agustín Piqueres, Andres Marafioti, Khalid Almubarak, Sourab Mangrulkar, Younes Belkada, and Leandro von Werra. Smollm2: When smol goes big – data-centric training of a small language model, 2025.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021.
- Igor A Bolshakov. A method of linguistic steganography based on collocationally-verified synonymy. In *International Workshop on Information Hiding*, pages 180–191. Springer, 2004.
- Jack T Brassil, Steven Low, Nicholas F Maxemchuk, and Lawrence O’Gorman. Electronic marking and identification techniques to discourage document copying. *IEEE Journal on Selected Areas in Communications*, 13(8):1495–1504, 1995.
- Ching-Yun Chang and Stephen Clark. Practical linguistic steganography using contextual synonym substitution and a novel vertex coding method. *Computational linguistics*, 40(2):403–448, 2014.
- Mark Chapman, George I Davida, and Marc Rennhard. A practical and effective approach to large-scale automated linguistic steganography. In *International Conference on Information Security*, pages 156–165. Springer, 2001.
- Mark Chen et al. Evaluating large language models trained on code. *arXiv*, 2021.
- Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. *Cryptology ePrint Archive*, 2023.
- Evan Crothers, Nathalie Japkowicz, and Herna Viktor. Machine generated text: A comprehensive survey of threat models and detection methods. *arXiv preprint arXiv:2210.07321*, 2022.
- 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.
- Google DeepMind. Gemma 3: Multimodal, multilingual, long context open models. Technical report, Google, 2025.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letak, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. Three bricks to consolidate watermarks for large language models. *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2023.
- Yu Fu, Deyi Xiong, and Yue Dong. Watermarking conditional text generation for ai detection: Unveiling challenges and a semantic-aware watermark remedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18003–18011, 2024.
- Eva Giboulot and Teddy Furon. Watermax: breaking the llm watermark detectability-robustness-quality trade-off. *arXiv preprint arXiv:2403.04808*, 2024.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.
- Abe Bohan 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. *arXiv preprint arXiv:2310.03991*, 2023.
- Abe Bohan Hou, Jingyu Zhang, Yichen Wang, Daniel Khashabi, and Tianxing He. k-semstamp: A clustering-based semantic watermark for detection of machine-generated text. *arXiv preprint arXiv:2402.11399*, 2024.

- Baihe Huang, Banghua Zhu, Hanlin Zhu, Jason D. Lee, Jiantao Jiao, and Michael I. Jordan. Towards optimal statistical watermarking, 2023.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Singh Chaplot Devendra, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- Nikola Jovanovi  , Robin Staab, Maximilian Baader, and Martin Vechev. Ward: Provable rag dataset inference via llm watermarks. *ICLR*, 2025.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023a.
- 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, 2023b.
- Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.
- Gregory Kang Ruey Lau, Xinyuan Niu, Hieu Dao, Jiangwei Chen, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. Waterfall: Framework for robust and scalable text watermarking. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.
- Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. Who wrote this code? watermarking for code generation. *arXiv preprint arXiv:2305.15060*, 2023.
- Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A semantic invariant robust watermark for large language models. *arXiv preprint arXiv:2310.06356*, 2023.
- Yepeng Liu and Yuheng Bu. Adaptive text watermark for large language models. *arXiv preprint arXiv:2401.13927*, 2024.
- Hasan Mesut Meral, B  lent Sankur, A Sumru   zsoy, Tunga G  ng  r, and Emre Sevin  . Natural language watermarking via morphosyntactic alterations. *Computer Speech & Language*, 23(1):107–125, 2009.
- OpenAI. ChatGPT: Optimizing language models for dialogue., 2022.
- Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, et al. Markllm: An open-source toolkit for llm watermarking. *arXiv preprint arXiv:2405.10051*, 2024.
- Julien Piet, Chawin Sitawarin, Vivian Fang, Norman Mu, and David Wagner. Mark my words: Analyzing and evaluating language model watermarks. *arXiv preprint arXiv:2312.00273*, 2023.
- Project Gutenberg. Project gutenberg, 2025. Accessed: 2025-11-15.
- Jipeng Qiang, Shiyu Zhu, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. Natural language watermarking via paraphraser-based lexical substitution. *Artificial Intelligence*, 317:103859, 2023.
- Wenjie Qu, Dong Yin, Zixin He, Wei Zou, Tianyang Tao, Jinyuan Jia, and Jiaheng Zhang. Provably robust multi-bit watermarking for ai-generated text via error correction code. *arXiv preprint arXiv:2401.16820*, 2024.
- Saksham Rastogi, Pratyush Maini, and Danish Pruthi. Stamp your content: Proving dataset membership via watermarked rephrasings. *arXiv preprint arXiv:2504.13416*, 2025.
- Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. Watermarking makes language models radioactive. *NeurIPS*, 2024.
- Tom Sander, Pierre Fernandez, Saeed Mahloujifar, Alain Durmus, and Chuan Guo. Detecting benchmark contamination through watermarking. *arXiv preprint arXiv:2502.17259*, 2025.
- M Hassan Shirali-Shahreza and Mohammad Shirali-Shahreza. A new synonym text steganography. In *2008 international conference on intelligent information hiding and multimedia signal processing*, pages 1524–1526. IEEE, 2008.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NeurIPS*, 2014.
- Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2409.12117*, 2024.

- Mercan Topkara, Cuneyt M Taskiran, and Edward J Delp III. Natural language watermarking. In *Security, Steganography, and Watermarking of Multimedia Contents VII*, pages 441–452. SPIE, 2005.
- Mercan Topkara, Giuseppe Riccardi, Dilek Hakkani-Tür, and Mikhail J Atallah. Natural language watermarking: Challenges in building a practical system. In *Security, Steganography, and Watermarking of Multimedia Contents VIII*, pages 106–117. SPIE, 2006a.
- Mercan Topkara, Umut Topkara, and Mikhail J Atallah. Words are not enough: sentence level natural language watermarking. In *Proceedings of the 4th ACM international workshop on Contents protection and security*, pages 37–46, 2006b.
- Umut Topkara, Mercan Topkara, and Mikhail J Atallah. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions. In *Proceedings of the 8th workshop on Multimedia and security*, pages 164–174, 2006c.
- Honai Ueoka, Yugo Murawaki, and Sadao Kurohashi. Frustratingly easy edit-based linguistic steganography with a masked language model. *arXiv preprint arXiv:2104.09833*, 2021.
- Ashish Venugopal, Jakob Uszkoreit, David Talbot, Franz Josef Och, and Juri Ganitkevitch. Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1363–1372, 2011.
- Zongqi Wang, Tianle Gu, Baoyuan Wu, and Yujiu Yang. Morphmark: Flexible adaptive watermarking for large language models. *arXiv preprint arXiv:2505.11541*, 2025.
- Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, 2022.
- Alex Wilson and Andrew D Ker. Avoiding detection on twitter: embedding strategies for linguistic steganography. *Electronic Imaging*, 28:1–9, 2016.
- Keith Winstein. Lexical steganography through adaptive modulation of the word choice hash. *Unpublished*. <http://www.imsa.edu/~keithw/tlex>, 1998.
- Yihan Wu, Zhengmian Hu, Hongyang Zhang, and Heng Huang. Dipmark: A stealthy, efficient and resilient watermark for large language models. *arXiv preprint arXiv:2310.07710*, 2023.
- Lingyun Xiang, Xinhui Wang, Chunfang Yang, and Peng Liu. A novel linguistic steganography based on synonym run-length encoding. *IEICE transactions on Information and Systems*, 100(2):313–322, 2017.
- Xiaojun Xu, Jinghan Jia, Yuanshun Yao, Yang Liu, and Hang Li. Robust multi-bit text watermark with llm-based paraphrasers. *arXiv preprint arXiv:2412.03123*, 2024.
- KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. Robust multi-bit natural language watermarking through invariant features. *arXiv preprint arXiv:2305.01904*, 2023a.
- KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. Advancing beyond identification: Multi-bit watermark for language models. *arXiv preprint arXiv:2308.00221*, 2023b.
- KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. Advancing beyond identification: Multi-bit watermark for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4031–4055, 2024.
- Jingqi Zhang, Ruibo Chen, Yingqing Yang, Peihua Mai, Heng Huang, and Yan Pang. Leave no trace: Black-box detection of copyrighted dataset usage in large language models via watermarking. *arXiv preprint arXiv:2510.02962*, 2025.
- Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Farinaz Koushanfar. {REMARK-LLM}: A robust and efficient watermarking framework for generative large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1813–1830, 2024.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.
- Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*, 2023.

Appendix

A Additional Details on LLM Watermarking

A.1 Hash Function

All watermarking schemes in this work rely on a pseudorandom function (PRF) to deterministically map how watermark windows should influence the next-token selection (e.g., green/red classification or Gumbel noise values). The PRF takes as input the candidate token x , a context window $\mathbf{w} = (w_1, \dots, w_k)$ of k token IDs, and the secret key s (all of them are integers), and outputs a random integer in $[0, M)$. We compute:

$$h'(x, \mathbf{w}, s) = \left(p_2 \cdot x + \sum_{i=1}^k w_i \cdot q_i + p_3 \cdot s \right) \cdot p_4, \quad (1)$$

$$h(x, \mathbf{w}, s) = \text{XORShift}(h'(x, \mathbf{w}, s)) \bmod M, \quad (2)$$

where q_1, \dots, q_k are distinct large primes (to ensure that different orderings of the same tokens produce different tokens), and p_2, p_3, p_4 are additional primes. The first result h' undergoes XOR-shift for better bit dispersion: $h = (h' \cdot p_{\text{mix}}) \oplus ((h' \cdot p_{\text{mix}}) \gg s)$, where p_{mix} is a mixing prime and s is a shift constant. Finally, $h \bmod M$ yields an integer in $[0, M)$.

For uniform PRF output in $[0, 1)$, we divide by M . For binary (green/red) classification, we threshold at γ : a token is “green” if $h/M < \gamma$. This construction ensures that small changes in any input (window, token, or key) produce uncorrelated outputs, satisfying the pseudo-randomness requirements for watermark security. In practice, the implementation allows for selectively scoring some tokens without needing to generate the full green/red list, which can speed up computations for large vocabularies, especially for SynthID-text, which uses multiple lists for different g -values.

A.2 Details on Watermark Schemes

We describe below the main watermarking schemes we evaluate.

Green-list/red-list. Kirchenbauer et al. (2023a,b) modify the logit vector during next-token generation based on the watermark window of k previous tokens and the private key s . Note that the case where $k = 0$ corresponds to the work of Zhao et al. (2023), but we found that it empirically breaks theoretical assumptions and often leads to degenerate text; therefore, we use $k \geq 1$ in our experiments. For each token $v \in \mathcal{V}$, the pseudorandom function $\text{PRF}(\cdot)$ outputs a value in $[0, 1)$ from the token ID v , the context window, and s . A token is classified as “green” if $\text{PRF}(v, \mathbf{w}, s) < \gamma$, where $\gamma \in [0, 1]$ is the expected proportion of green tokens under the null hypothesis (typically $\gamma = 0.5$). Logits of green tokens are incremented by δ to increase their sampling probability:

$$\tilde{\ell}_v = \begin{cases} \ell_v + \delta & \text{if } v \in \text{GreenList} \\ \ell_v & \text{otherwise} \end{cases} \quad (3)$$

Detection involves repeating the greenlist computation for each token of a text, incrementing a score by 1 if the token is in the greenlist, and performing a statistical test on the cumulative score. Under the null hypothesis \mathcal{H}_0 “the text is not watermarked with that scheme and private key s ”, this score follows a binomial distribution. A simple binomial test thus provides a p -value: the probability of observing at least as many green tokens as observed, if the text was not watermarked.

SWEET. Lee et al. (2023) apply the Green-red list watermark only to high-entropy tokens. The intuition is that low-entropy tokens (where the model is confident) contribute little signal but can degrade quality if biased. At generation time, tokens with entropy below a threshold are sampled without watermark bias. Specifically, a token is considered high-entropy if $H(\mathbf{p}) = -\sum_{v \in \mathcal{V}} p_v \log p_v > \tau$, where τ is a predefined threshold. At detection time, these tokens are similarly excluded from scoring, improving signal-to-noise ratio. Note that this filtering can also be applied only at detection time, and on top of other schemes than green-red (as we do in our experiments).

MorphMark. Wang et al. (2025) adaptively adjust watermark strength based on context. Let $P_G = \sum_{v \in \text{GreenList}} p_v$ be the total probability mass on green tokens before watermarking. If $P_G \leq p_0$ (a threshold, e.g., $p_0 = 0.15$), no watermark is applied to preserve quality. Otherwise, an adaptive boost factor $r = \min(\kappa P_G, 1)$ is computed, where κ controls the watermark strength. Probabilities are then adjusted:

$$\hat{p}_v = \begin{cases} p_v \cdot \left(1 + \frac{r(1-P_G)}{P_G}\right) & \text{if } v \in \text{GreenList} \\ p_v \cdot (1 - r) & \text{otherwise} \end{cases} \quad (4)$$

This ensures stronger watermarking when the green list is already favorable, minimizing quality degradation.

DiPMark. Wu et al. (2023) introduce a variant of Green-red watermarks that is distortion-free. The method uses a pseudorandom permutation π (seeded by the context window and s) to reorder tokens. After permutation, the cumulative probability distribution is modified as follows: tokens in the interval $[0, 1 - \alpha]$ have their probability set to zero, tokens in $[1 - \alpha, \alpha]$ remain unchanged, and tokens in $[\alpha, 1]$ have their probability doubled (then renormalized). This creates a detectable bias while preserving the original distribution’s average over the randomness of π .

Gumbel-max. Aaronson and Kirchner (2023) alternatively leverage the “Gumbel-max trick”. After applying temperature scaling and optional top- k or top- p filtering to obtain a probability vector \mathbf{p} , the next token is selected as:

$$x_t = \arg \max_{v \in \mathcal{V}} r_v^{1/p_v} \quad (5)$$

where $r_v \sim \text{Uniform}(0, 1)$ is i.i.d. noise (which is equivalent to sampling from \mathbf{p}). The watermark intervenes by replacing the purely random r_v with pseudorandom values $r_v = \text{PRF}(v, \mathbf{w}, s) \in [0, 1]$ generated by the PRF from the token ID v , the context window, and s , for each $v \in \mathcal{V}$. Consequently, for a fixed context, the noise vector is deterministic. Detection is performed by recomputing the PRF output $r_{x_t} = \text{PRF}(x_t, \mathbf{w}_t, s)$ for each observed token x_t and computing the cumulative score:

$$S = \sum_{t=1}^N -\log(1 - r_{x_t}) \quad (6)$$

where N is the number of tokens in the text. Under \mathcal{H}_0 “the text is not watermarked with that scheme and private key s ”, S follows a $\Gamma(N, 1)$ distribution. The p -value of a test associated with score S reads: $p\text{-value}(S) = \Gamma(N, S)/\Gamma(N)$, where Γ is the upper incomplete gamma function.

SynthID-text. Dathathri et al. (2024) propose SynthID Text, which employs a tournament-based sampling strategy. The method uses m scoring functions g_1, \dots, g_m , where each g_i maps a token to a pseudorandom value in $\{0, 1\}$ based on the context window and secret key¹. Unlike the additive bias of the “Green-red” list method, this approach samples 2^m candidate tokens from the model’s distribution and randomly organizes them into a tournament where winners are determined by the g -values. Detection relies on a statistical hypothesis test using the mean g -value of the observed tokens as the test statistic. Under the null hypothesis \mathcal{H}_0 (unwatermarked text), the g -values are independent and identically distributed (e.g., uniform or Bernoulli). This scheme is shown to outperform Green-red approaches because the tournament depth m allows for a more favorable trade-off between detectability and text quality (perplexity), particularly in low-entropy settings where standard additive biases often degrade coherence.

Scoring Functions: Mean vs. Weighted Mean. While the standard detection method computes a simple unweighted mean of the g -values across all tokens and tournament layers (effectively a normalized sum analogous to a binomial test), Dathathri et al. (2024) demonstrate that this is theoretically suboptimal for multi-layer tournaments. The authors observe that the amount of “watermarking evidence” embedded by the tournament is not uniform across layers; specifically, the strength of the watermark diminishes as the tournament depth increases because each successive layer operates on a subset of candidates with reduced entropy. Consequently, an unweighted mean dilutes the strong signal

¹In our implementation, each g -function corresponds to a distinct green/red list partition: we compute them by incrementing the token ID by i for each depth i of the tournament. Since the hash function exhibits no correlation for incremented token IDs, this effectively simulates independent green/red lists per tournament layer.

Table 6 Key sensitivity analysis across models and watermarking schemes. We evaluate multiple random keys per configuration on 100 non-watermarked texts. “SynthID (W)” denotes the Weighted variant which utilizes a different scoring statistic. The **Best Key** is selected based on the highest p -value from the Kolmogorov-Smirnov (KS) test for uniformity under \mathcal{H}_0 .

Model	Method	Aggregated Statistics (All Keys)		Selected “Best” Key		
		Avg. p -val	σ_{keys}	Key ID	Avg. p -val	KS p
Qwen 2.5	DipMark	0.501	0.033	#596061	0.493	0.99
	Green-red	0.501	0.033	#596061	0.493	0.99
	MorphMark	0.501	0.033	#596061	0.493	0.99
	Gumbel-max	0.497	0.031	#2345	0.499	0.99
	SynthID ($d = 10$)	0.498	0.035	#606	0.494	0.95
	SynthID ($d = 20$)	0.488	0.033	#753	0.506	0.99
	SynthID ($d = 30$)	0.488	0.031	#1357	0.493	0.97
	SynthID (W) ($d = 10$)	0.498	0.030	#323334	0.506	0.98
	SynthID (W) ($d = 20$)	0.493	0.029	#3452	0.498	0.99
	SynthID (W) ($d = 30$)	0.490	0.032	#505152	0.509	1.00
Llama 3	DipMark	0.508	0.037	#323334	0.505	0.95
	Green-red	0.508	0.037	#323334	0.505	0.95
	MorphMark	0.508	0.037	#323334	0.505	0.95
	Gumbel-max	0.498	0.030	#6780	0.506	1.00
	SynthID ($d = 10$)	0.507	0.035	#656667	0.504	0.99
	SynthID ($d = 20$)	0.497	0.035	#626364	0.493	0.99
	SynthID ($d = 30$)	0.503	0.037	#686970	0.499	0.99
	SynthID (W) ($d = 10$)	0.504	0.036	#888	0.507	0.96
	SynthID (W) ($d = 20$)	0.500	0.034	#42	0.504	1.00
	SynthID (W) ($d = 30$)	0.502	0.036	#333	0.490	0.98
Gemma 3	DipMark	0.501	0.035	#8907	0.488	0.96
	Green-red	0.501	0.035	#8907	0.488	0.96
	MorphMark	0.501	0.035	#8907	0.488	0.96
	Gumbel-max	0.498	0.034	#6789	0.506	0.97
	SynthID ($d = 10$)	0.503	0.041	#656667	0.504	0.98
	SynthID ($d = 20$)	0.496	0.036	#323334	0.488	0.99
	SynthID ($d = 30$)	0.502	0.036	#131415	0.495	0.93
	SynthID (W) ($d = 10$)	0.501	0.042	#1234	0.519	0.97
	SynthID (W) ($d = 20$)	0.492	0.036	#258	0.501	0.99
	SynthID (W) ($d = 30$)	0.503	0.039	#222324	0.498	0.98
SmolLM2	DipMark	0.496	0.034	#789	0.506	1.00
	Green-red	0.496	0.034	#789	0.506	1.00
	MorphMark	0.496	0.034	#789	0.506	1.00
	Gumbel-max	0.507	0.038	#252627	0.498	1.00
	SynthID ($d = 10$)	0.506	0.044	#535455	0.501	0.98
	SynthID ($d = 20$)	0.495	0.041	#369	0.499	1.00
	SynthID ($d = 30$)	0.497	0.043	#444	0.500	0.99
	SynthID (W) ($d = 10$)	0.502	0.043	#707	0.509	0.98
	SynthID (W) ($d = 20$)	0.499	0.041	#369	0.499	1.00
	SynthID (W) ($d = 30$)	0.501	0.044	#252627	0.507	0.98

from early layers with the weaker signal from deeper layers. To address this, SynthID-text utilizes a *weighted mean score*, which assigns decreasing weights w_l to the g -values of the l -th tournament layer. By emphasizing the earlier layers where the statistical signature is most robust, the weighted scoring function improves the signal-to-noise ratio of the test statistic, yielding higher detection accuracy (true positive rate) for a fixed false positive rate compared to the classical unweighted approach.

A.3 Choice of the Secret Key and Statistical Correctness

How to choose the secret key? As explained in [section 3](#), valid detection requires p -values to be uniformly distributed under \mathcal{H}_0 . While this holds in expectation over all possible keys, fixing a specific key s creates preferences for certain $(k + 1)$ -grams. If these patterns align with natural language statistics, the expected green fraction under \mathcal{H}_0 shifts slightly (e.g., from 0.5 to 0.505), causing p -values to collapse toward zero on long texts and inflating false positive rates.

We address this by testing 50 candidate keys for each combination of tokenizer, watermarking method, and relevant hyperparameters. Each key is evaluated on 100 non-watermarked texts of 800 characters, and we select the key that minimizes deviation from $U(0, 1)$ as measured by the Kolmogorov-Smirnov (KS) statistic. This ensures our results reflect intrinsic scheme performance rather than key-specific artifacts. Detailed statistics are provided in [Table 6](#).

A.4 Details on Radioactivity

The Radioactivity Test Protocol. To formally test for watermarked dataset radioactivity, we detail the methodology from Sander et al. (2024, 2025) for the Green-list/Red-list scheme. The core idea is to repurpose the standard watermark detection test (normally applied to observed text tokens) and instead apply it to the *predicted* tokens of the suspect model. This allows us to determine whether the model is familiar with the watermark, thereby providing evidence that the model was exposed to the watermark during training.

Teacher Forcing Setup. We feed the *watermarked* text \mathbf{y} into the suspect model f_θ using a teacher-forcing setup. Let \hat{y}_t denote the top-1 prediction of the suspect model at step t , given the watermarked prefix $y_{<t}$:

$$\hat{y}_t = \operatorname{argmax}_{v \in \mathcal{V}} P_\theta(v \mid y_{<t}). \quad (7)$$

Test Statistic. We define the radioactivity score S as the empirical proportion of the suspect model’s predictions that fall into the Green-list:

$$S = \frac{1}{|\mathcal{U}|} \sum_{t \in \mathcal{U}} \mathbb{I}(h(\hat{y}_t, y_{t-k:t-1}, s) < \gamma), \quad (8)$$

where s is the secret key and γ is the expected green ratio (typically 0.5).

Deduplication. To ensure statistical independence, we must account for repeated n-grams in the context that could influence the model to reproduce them. In particular, if an n-gram in context is green, the suspect model might repeat it due to context copying rather than watermark radioactivity. We filter the indices t to form a set \mathcal{U} such that each watermark window $y_{t-k:t-1}$ is only scored once, which fully removes this issue (Sander et al., 2024). Under the null hypothesis \mathcal{H}_0 (i.e., the suspect model is unaware of s), the count of green predictions should now follow a binomial distribution, as the suspect model should exhibit no preference toward green or red tokens:

$$K \sim \text{Binomial}(|\mathcal{U}|, \gamma). \quad (9)$$

This formulation allows for the computation of an exact p -value.

WaterMax Is Not Radioactive. For the score to truly follow a known binomial distribution under \mathcal{H}_0 (“the suspect model has never been in contact with the watermark”), the watermark bias must be applied at the *token level*. However, WaterMax selects the final sequence x^* from a set of candidates \mathcal{C} by maximizing the number of green tokens globally *by chance*. Consequently, an innocent model f_θ that shares a similar language distribution with the generator P_{gen} will find that its optimal next token \hat{y}_t is green significantly more often than γ , purely due to this selection bias.

We verified this experimentally by running the radioactivity detection test on approximately 100k watermarked tokens generated via WaterMax, Green-list/Red-list random sampling, and Green-list/Red-list beam search with biased scoring (after rephrasing 300 excerpts from Dickens as detailed in subsection 4.1). As expected, the radioactivity test yields a p -value of 0.93 for Green-list/Red-list random sampling and 0.76 for Green-list/Red-list beam search with biased scoring, *but* 1.0×10^{-6} for *WaterMax*.

B Additional Text Samples

B.1 Example of Code Tasks and Watermarked Codes

Input (Prompt + Canonical Solution):

```
from typing import List

def has_close_elements(numbers: List[float], threshold:
    float) -> bool:
    """ Check if in given list of numbers, are any two
        numbers closer to each other than given threshold.
    >>> has_close_elements([1.0, 2.0, 3.0], 0.5)
    False
    >>> has_close_elements([1.0, 2.0, 5.0, 2.0], 0.3)
    True
    """
```

```
for idx, elem in enumerate(numbers):
    for idx2, elem2 in enumerate(numbers):
        if idx != idx2:
            distance = abs(elem - elem2)
            if distance < threshold:
                return True
    return False
```

Test:

```
METADATA = { 'author': 'jt', 'dataset': 'test' }

def check(candidate):
    assert candidate([1.0, 2.0, 3.9, 4.0, 5.0, 2], 0.3) == True
    assert candidate([1.0, 2, 3, 4.0, 5.0, 2.2], 0.05) == False
    assert candidate([1.0, 2.0, 5.9, 4.0, 5.0, 0.95]) == True
    assert candidate([1.0, 2.0, 5.9, 4.0, 5.0, 0.8]) == False
    assert candidate([1.0, 2.0, 3.0, 4.0, 5.0, 2], 0.1) == True

check(has_close_elements)
```

Figure 8 Example of a HumanEval task. The left block represents the text subjected to post-hoc watermarking (paraphrasing). The right block is the test harness used to verify functional correctness (Pass@1).

Table 7 Post-hoc code watermarking examples on a sample from MBPP (using Llama-3.1-8B at temperatures 0.8, 1.2 and 1.4 with Gumbel-max watermarking, top- $p = 0.95$).

Original Code	Watermarked Code
<pre># A function to find the minimum cost path to reach (m, n) from (0, 0) ... R = 3 C = 3 def min_cost(cost, m, n): tc = [[0 for x in range(C)] for x in range(R)] tc[0][0] = cost[0][0] for i in range(1, m+1): tc[i][0] = tc[i-1][0] + cost[i][0] for j in range(1, n+1): tc[0][j] = tc[0][j-1] + cost[0][j] for i in range(1, m+1): for j in range(1, n+1): tc[i][j] = min(tc[i-1][j-1], tc[i-1][j], tc[i][j-1]) + cost[i][j] return tc[m][n]</pre>	<pre># A function to calculate the minimum cost path from the origin (0, 0) ... R = 3 C = 3 def min_cost_path(cost, m, n): # Create a temporary cost matrix tc of the same dimensions... tc = [[0 for x in range(C)] for x in range(R)] # The cost to reach the first cell (0, 0) is the cost... tc[0][0] = cost[0][0] # Calculate the cost to reach the first row. for i in range(1, m+1): # The cost to reach the first cell in the current row... tc[i][0] = tc[i-1][0] + cost[i][0] # Calculate the cost to reach the first column. for j in range(1, n+1): # The cost to reach the first cell in the current column... tc[0][j] = tc[0][j-1] + cost[0][j] # Calculate the cost to reach the rest of the cells. for i in range(1, m+1): for j in range(1, n+1): # The cost to reach the current cell is the minimum... tc[i][j] = min(tc[i-1][j-1], tc[i-1][j], tc[i][j-1]) + cost[i][j] # Return the cost to reach the cell at position (m, n). return tc[m][n]</pre>
<p>Stats: T: 0.8 • p_{value}: $2.11e-01$ • Tokens: 414 • Code correctness: True</p>	

Table 7 Post-hoc code watermarking at different temperatures (continued)

Original Code	Watermarked Code
<pre> # A function to find the minimum cost path to reach (m, n) from (0, 0) ... R = 3 C = 3 def min_cost(cost, m, n): tc = [[0 for x in range(C)] for x in range(R)] tc[0][0] = cost[0][0] for i in range(1, m+1): tc[i][0] = tc[i-1][0] + cost[i][0] for j in range(1, n+1): tc[0][j] = tc[0][j-1] + cost[0][j] for i in range(1, m+1): for j in range(1, n+1): tc[i][j] = min(tc[i-1][j-1], tc[i-1][j], tc[i][j-1]) + cost[i][j] return tc[m][n] </pre>	<pre> # Function to calculate the minimum cost path from the origin (0, 0) ... R, C = 3, 3 def min_cost_path(cost, m, n): dp = [[0 for _ in range(C)] for _ in range(R)] dp[0][0] = cost[0][0] # Calculate the minimum cost to reach the first row. for i in range(1, m+1): dp[i][0] = dp[i-1][0] + cost[i][0] # Calculate the minimum cost to reach the first column. for j in range(1, n+1): dp[0][j] = dp[0][j-1] + cost[0][j] # Calculate the minimum cost to reach the rest of the positions. for i in range(1, m+1): for j in range(1, n+1): dp[i][j] = min(dp[i-1][j-1], dp[i-1][j], dp[i][j-1]) + cost[i][j] return dp[m][n] </pre>
Stats: T : 1.2 • p_{value} : $2.63e-02$ • Tokens: 256 • Code correctness: True	
<pre> # A function to find the minimum cost path to reach (m, n) from (0, 0) ... R = 3 C = 3 def min_cost(cost, m, n): tc = [[0 for x in range(C)] for x in range(R)] tc[0][0] = cost[0][0] for i in range(1, m+1): tc[i][0] = tc[i-1][0] + cost[i][0] for j in range(1, n+1): tc[0][j] = tc[0][j-1] + cost[0][j] for i in range(1, m+1): for j in range(1, n+1): tc[i][j] = min(tc[i-1][j-1], tc[i-1][j], tc[i][j-1]) + cost[i][j] return tc[m][n] </pre>	<pre> # Function to calculate the minimum cost path from the origin (0, 0) ... RatingsColumns = 3 Rows = 3 def min_expense(cost, m, n): databaseOfDiscountedValues = [[0 for xPos in range(RatingsColumns)] for yPos in range(Rows)] initialPositionValuation = cost[0][0] for tabIndex in range(1, Rows+1): databaseOfDiscountedValues[tabIndex][0] = databaseOfDiscountedValues[tabIndex-1][0] + cost[tabIndex][0] for tabIndex in range(1, RatingsColumns+1): databaseOfDiscountedValues[0][tabIndex] = databaseOfDiscountedValues[0][tabIndex-1] + cost[0][tabIndex] for tabIndex in range(1, Rows+1): for colIndex in range(1, RatingsColumns+1): dbDV = min(databaseOfDiscountedValues[tabIndex-1][colIndex-1], databaseOfDiscountedValues[tabIndex-1][colIndex], databaseOfDiscountedValues[tabIndex][colIndex-1]) + cost[tabIndex][colIndex] databaseOfDiscountedValues[tabIndex][colIndex] = dbDV return databaseOfDiscountedValues[m][n] </pre>
Stats: T : 1.4 • p_{value} : $8.14e-11$ • Tokens: 302 • Code correctness: False	

B.2 Descriptive Text and Dialogue Preservation

Table 8 Additional qualitative comparisons on literary text. Additional samples from *A Christmas Carol* by Charles Dickens, rephrased using Llama-3.2-3B-Instruct with the Gumbel-max scheme, $T = 0.7$ and top-p = 0.95. All outputs maintain strong semantic similarity and watermark detectability.

Original Text	Watermarked Rephrasing
<i>Example 1</i>	
Lumber-room as usual. Old fire-guard, old shoes, two fish-baskets, washing-stand on three legs, and a poker. Quite satisfied, he closed his door, and locked himself in; double-locked himself in, which was not his custom. Thus secured against surprise, he took off his cravat; put on his dressing-gown and slippers, and his nightcap; and sat down before the fire to take his gruel.	Lumber-room as usual. Old fire-guard, worn shoes, two fish-baskets, a washing-stand on three legs, and a poker. He was quite content, so he locked himself in, double-locked himself in, which wasn't his usual practice. Thus secure against surprise, he shed his cravat, put on his dressing-gown, slippers, and his nightcap; and sat down before the fire to enjoy his gruel.
It was a very low fire indeed; nothing on such a bitter night. He was obliged to sit close to it, and brood over it, before he could extract the least sensation of warmth from such a handful of fuel. The fireplace was an old one, built by some Dutch merchant long ago, and paved all round with quaint Dutch tiles, designed to illustrate the Scriptures. There were Cains and Abels, Pharaoh's daughters; Queens of Sheba, Angelic messengers descending through the air on clouds like feather-beds, Abrahams, Belshazzars, Apostles putting off to sea in butter-boats, hundreds of figures to attract his thoughts; and yet that face of Marley, seven years dead, came like the ancient Prophet's rod, and swallowed up the whole.	It was a very low fire indeed; the flames seemed to be burning with a fierce intensity. He was forced to sit close to it, and reflect on it, before he could feel any warmth emanating from such a small amount of fuel. The fireplace was an old one, built by a Dutch merchant many years ago, and adorned with intricate tiles depicting biblical scenes. There were numerous figures from the Bible, including Cain and Abel, Pharaoh's daughters, the Queen of Sheba, and the Angelic messengers descending from the heavens like angels on cloud-like bodies; yet, his gaze fell upon the face of Marley, seven years deceased, and it seemed to hold a profound power, as if the very spirit of the departed was still present.
Stats: SBERT: 0.925 • Ratio: 1.03 • Perplexity: 3.10 • p -val: 6.7×10^{-4}	
<i>Example 2</i>	
"again; and followed it up with "Humbug." "Don't be cross, uncle!" said the nephew. "What else can I be," returned the uncle, "when I live in such a world of fools as this? Merry Christmas! Out upon merry Christmas! What's Christmas time to you but a time for paying bills without money; a time for finding yourself a year older, but not an hour richer; a time for balancing your books and having every item in 'em through a round dozen of months presented dead against you? If I could work my will," said Scrooge indignantly, "every idiot who goes about with 'Merry Christmas' on his lips, should be boiled with his own pudding, and buried with a stake of holly through his heart. He should!" "Uncle!" pleaded the nephew. "Nephew!" returned the uncle sternly, "keep Christmas in your own way, and let me keep it in mine.	"Again; and it followed with an exclamation, "Humbug!" "Don't be cross, uncle!" said the nephew. "What else can I be," returned the uncle, "when I live in such a world of fools as this? Christmas is a time when everyone is trying to deceive themselves, pretending to be merry and happy, but in reality, it's just a chore to pay bills and get older without earning more. It's a time when people make lists and check them twice, but never actually achieve anything. If I could change my ways," said Scrooge indignantly, "every fool who goes about spreading 'Merry Christmas' on their lips should be shunned and left to rot. They should be boiled in their own misery, and buried with a curse of isolation through the desolate winter months." "Uncle!" pleaded the nephew. "Nephew!" returned the uncle sternly, "keep Christmas in your own way, and let me keep it in mine."
Stats: SBERT: 0.935 • Ratio: 1.04 • Perplexity: 3.28 • p -val: 2.7×10^{-5}	

C Additional Results

C.1 Analysis of Output Length Distributions

In the main text, we apply a filtering criterion to retain only rephrasings where the output length remains within a factor of 0.75 to 1.25 of the original input length. In this appendix, we analyze the distribution of these length ratios to validate that this filtering does not introduce bias against specific watermarking schemes.

Figure 9 presents the distribution of length ratios (defined as $L_{\text{output}}/L_{\text{input}}$) across all evaluated models and five distinct watermarking methods. The green shaded region indicates the acceptance window used in our main experiments.

We observe two key trends:

1. **Consistency across schemes:** For any given base model (rows), the distribution of output lengths is remarkably consistent across all watermarking methods (columns). Whether using simple rejection sampling (WaterMax) or distribution perturbation (DiPMark, Gumbel-max), the variance in output length remains stable.
2. **Dependence on model capability:** The ability to respect the length constraint is primarily a function of the model size and instruction-following capability. Smaller models (e.g., SmolLM2-135M) exhibit high variance and frequently generate outputs that are too short or too long, whereas larger, more capable models (e.g., Llama-3.1-8B, Gemma-2-27B) produce tight distributions centered near the ideal ratio of 1.0.

These findings confirm that outliers in length are attributable to the underlying model’s generation stability rather than the watermarking process itself. Consequently, filtering these outliers allows for a fairer assessment of semantic preservation and detection power on valid generations, rather than penalizing watermarking schemes for the base model’s verbosity or brevity failures.

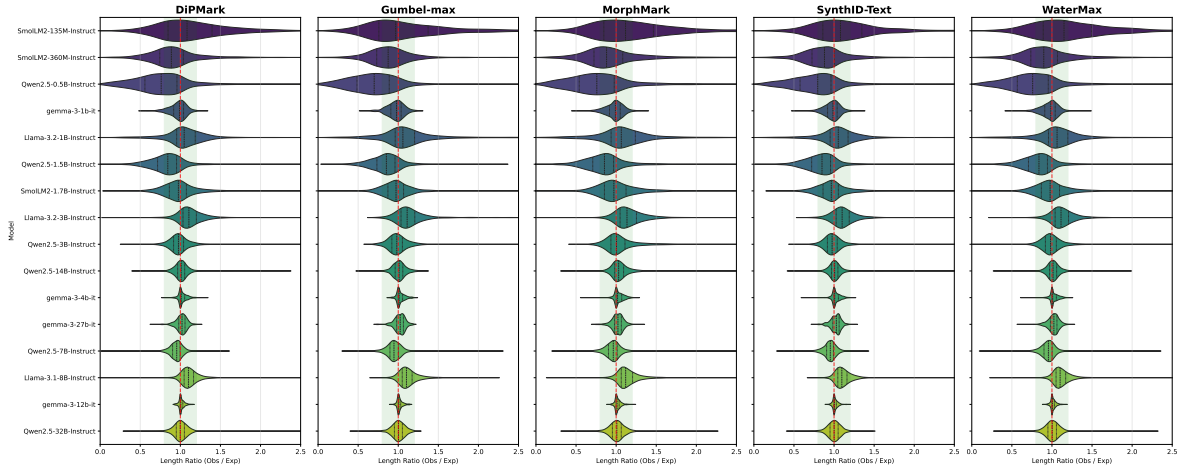


Figure 9 Impact of Model Choice on Output Length Consistency. Violin plots showing the distribution of length ratios ($\text{Length}_{\text{wm}}/\text{Length}_{\text{orig}}$) for varying models and watermarking schemes. The green shaded region represents the inclusion criteria ($[0.75, 1.25]$) used in the main experiments. The red dashed line indicates the ideal ratio of 1.0. Note that length variance is driven primarily by the base model choice (rows) rather than the watermarking scheme (columns), with larger models consistently adhering closer to the target length.