A Multi-bit Watermarking Scheme for LLM-Generated Short Texts

Jarosław Janas [1] Paweł Morawiecki [1] Josef Pieprzyk [1,2]

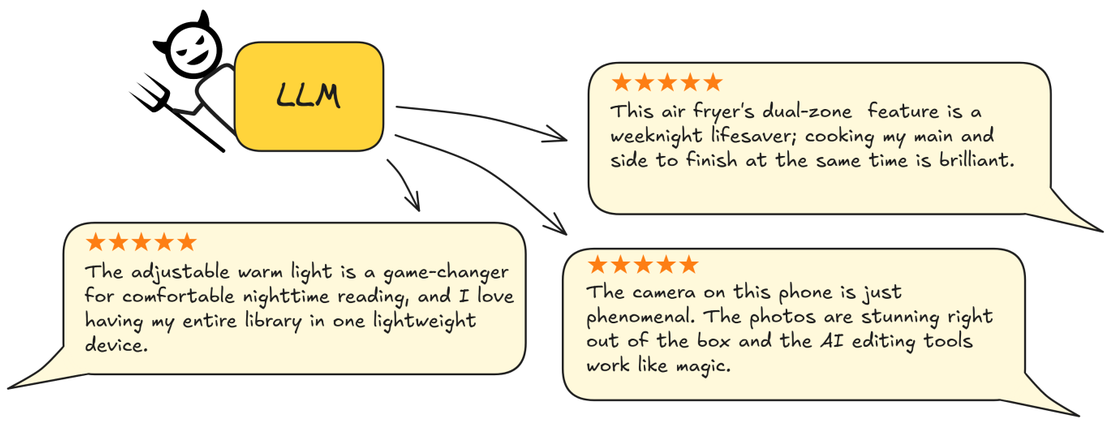
[1]Institute of Computer Science, Polish Academy of Sciences
[2]Data61, CSIRO, Australia

Abstract

We present a practical multi-bit watermarking scheme tailored for short (~100 words) texts (e.g., product reviews, social posts) generated by large language models (LLMs). The method uses a precomputed secret codebook that maps a n-bit watermark space to a longer u-bit codeword space. During generation, we bias token choice to encode the selected codeword; at detection, we recover a noisy u-bit sequence and perform a nearest-codeword lookup (Hamming distance). Key advantages: high match rates for 24-bit watermarks on short texts, simple extraction, and strong substitution robustness. Limitations: The extraction cost grows with codebook size, and the scheme is fragile to insertion/deletion (desynchronization) attacks.

Motivation

The ability of LLMs to generate vast quantities of convincing short texts, such as fake reviews, is a growing concern. As such, short-text LLM watermarking is essential. However, this poses a significant challenge, as the limited length of these texts provides little space to embed a message.



Contributions

- A codebook-based scheme for multi-bit watermarking that provides error-correction capabilities for short texts by mapping messages to longer, redundant codewords.
- An extraction method with manageable computational cost (for 24-bit watermark), based on a nearest-codeword search in Hamming space.

Method Overview

1. **Codebook Generation:** A secret codebook $\mathcal{C}$ is created by mapping each watermark $W \in \{0, \ldots, 2^n - 1\}$ to a unique, pseudorandom u-bit codeword, which is computed as $L_W = \mathsf{PRG}(H(W \,\|\, \mathcal{K}), u)$, where $\mathcal{K}$ is a secret key and H is a cryptographic hash function.
2. Select a watermark W and its corresponding codeword L_W.
3. **Embedding**: For each of the first u tokens, the vocabulary is pseudorandomly partitioned into two sets, $\mathcal{V}_0$ and $\mathcal{V}_1$, seeded by the previous token and key $\mathcal{K}$. Model logits are then biased to favour the set corresponding to the next bit of the codeword L_W.
4. **Extraction:** Iterate over the generated tokens, regenerating the same $\mathcal{V}_0$ and $\mathcal{V}_1$ partitions at each step to reconstruct a (potentially noisy) codeword candidate $\hat{L}$. Find the nearest true codeword L_W in $\mathcal{C}$ by minimum Hamming distance; accept if the distance is $\leq \eta$.

Experimental Setup

- **Parameters:** We embed a 24-bit watermark ($n = 24$) into a fixed-length text of 128 tokens ($u = 128$) using a logit bias of $\delta = 6$ and greedy sampling.
- **Dataset:** 400 prompts were used per model, sourced from the `HC3` and `essays-with-instructions` datasets. Only generations that successfully reached the full 128-token length were evaluated.
- **Evaluation Metrics:** We report the True Positive Rate (TPR) at various target False Positive Rates (FPR). Each FPR target defines a specific Hamming Distance (HD) threshold, determined by testing a large set of random bit sequences against the codebook. The TPR is then the fraction of valid generations whose extracted HD is at or below that threshold.

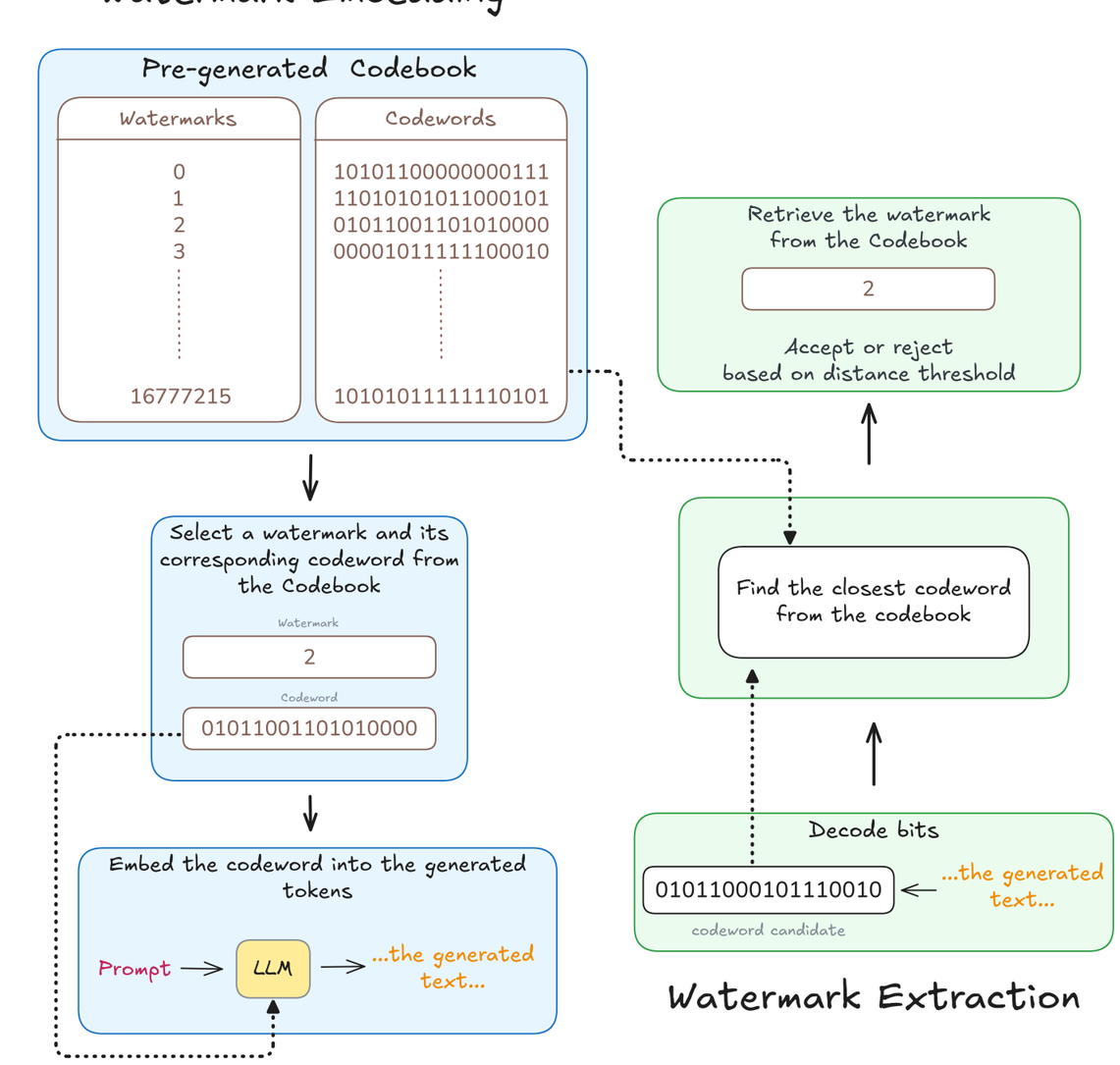


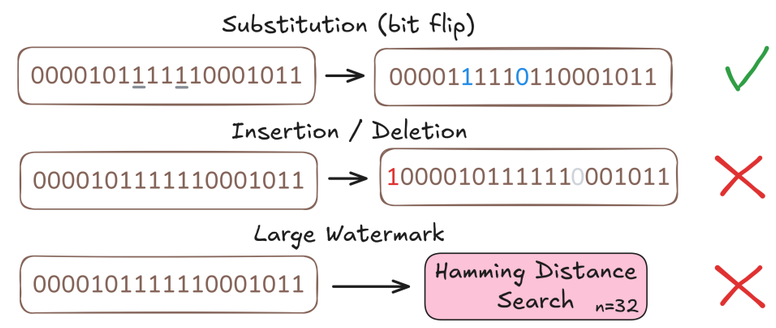
Figure 1. Method schematic: offline codebook, biased embedding, and nearest-codeword extraction.

Results: TPR at Various FPRs (codebook $n = 24$, $u = 128$)

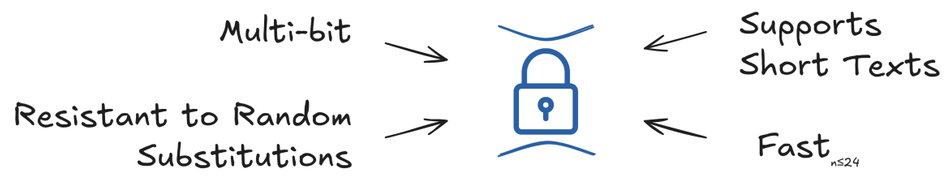
	FPR				
Model	1e-4	1e-3	1e-2	1e-1	1e0
HD Threshold	26	28	30	32	37
Llama 2 7B	98.4	98.7	99.1	99.1	99.1
Llama 3.1 8B	99.1	99.1	99.4	99.7	99.7
Llama 3.2 3B	100	100	100	100	100
Llama 3.2 3B (Instruct)	78.5	86.5	92.2	95.6	96.3
Mistral 7B	100	100	100	100	100
Mistral 7B (Instruct)	90.4	95.5	96.4	97.2	98.4

These results demonstrate high recovery rates on short outputs for a 24-bit watermark when using a 128-bit codeword. Note that *instruct* models tend to be more challenging.

Limitations



Conclusions



- We demonstrated a practical scheme that effectively embeds multi-bit watermarks (e.g., 24 bits) into short LLM-generated texts, achieving high recovery rates across multiple models.
- The core innovation - a codebook that converts extraction into a computationally manageable nearest-codeword search and is robust against substitution attacks and bit-flip errors.

This work was supported by the National Science Centre, Poland, under the grant 2023/49/B/ST6/02580