Practical and Provably Secure LLM Watermarking

Haw-Shiuan Chang[1], Varun Gandhi[1], Alex Hoover[2], **Meenakshi Iyer[1]**, Adam O'Neill[1]
University of Massachusetts, Amherst
University of Chicago

1. Introduction

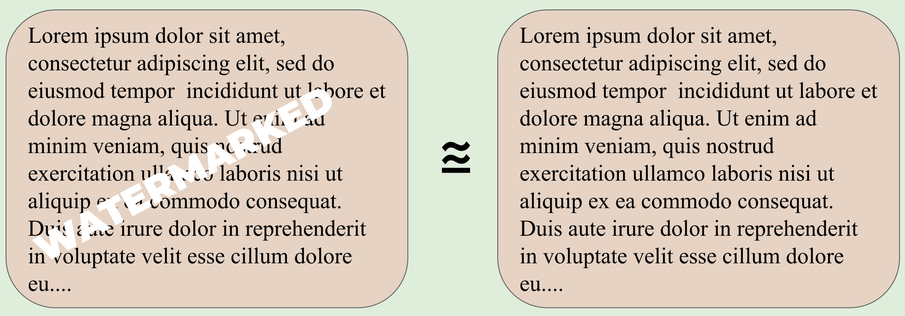
- LLMs have immense capability, but their usage can pose risks in public safety.
- Standard LLM-output detection models suffer from a high rate of false positives.
- Watermarking is a promising solution, offering a way to embed a detectable signal in text.

2. Goals

- Design a scheme that is robust, undetectable, and unforgeable.
- Combine theory (pseudorandom codewords) with practice (semantic robustness).

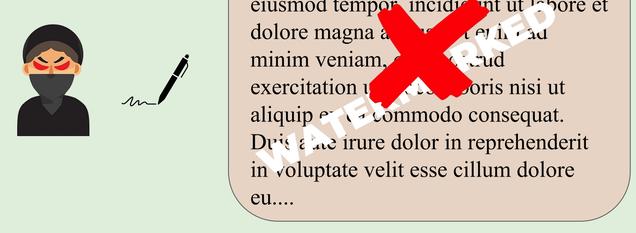
3. Watermarking Security

- Undetectability:

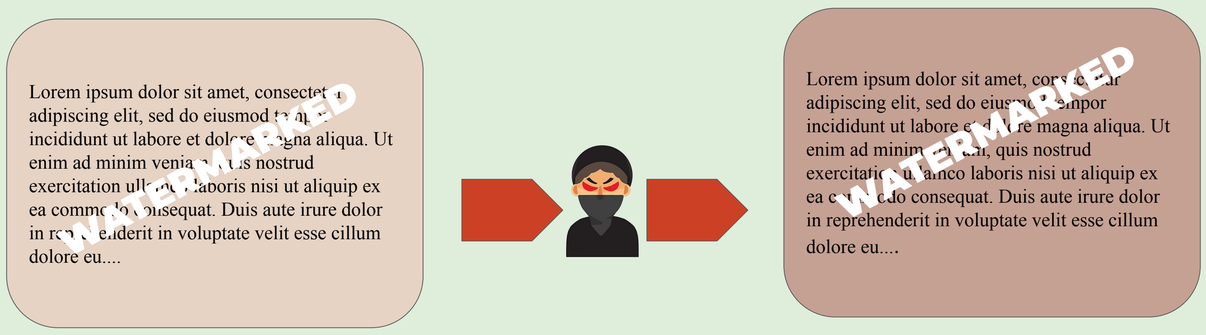


Watermarked Text ≅ Unwatermarked Text

- Unforgeability:



- Semantic Robustness:



4. Scheme Overview

Watermarking:

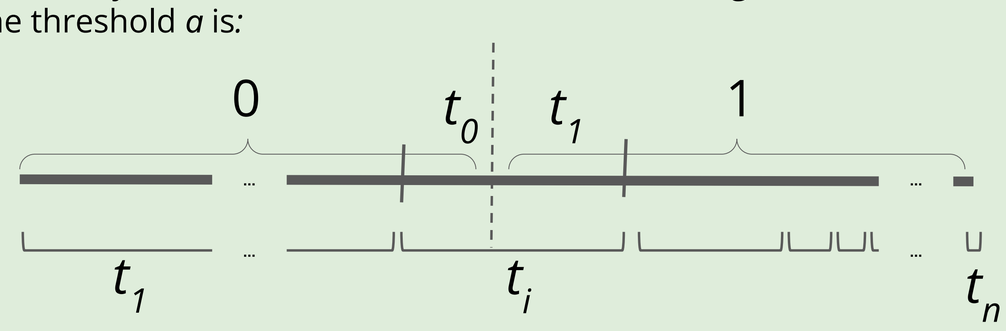
- Generate a codeword.
- Hash the previous 3 tokens generated to get a codeword index.
- Sample tokens from target-bit set, split at threshold α, starts at 0.5.
- If token not crossing threshold, repeat steps above.
- Else, sample next token with modified threshold.
- Continue till all text is generated.

Detection:

- Hash the previous 3 tokens generated to get a codeword index.
- Look at probability distribution and tokens sampled to determine bit embedded.
- Repeat until codeword recovered, verify codeword for watermark presence.

5. Threshold Determination

Probability distribution over tokens, sorted from highest to lowest.
The threshold α is:



$$\alpha = \frac{Pr(t_0)}{Pr(t_0) + Pr(t_1)}$$

6. Future Directions

- Generate empirical results. Does this scheme withstand a high rate of paraphrase attacks? Does the quality remain under LLM-as-a-judge?
- Analyze watermarking radioactivity. Does this watermark remain on LLMs that are trained on this text?
- Consider Private LLM Watermarking. Is it possible to watermark text without seeing it?

References

[CG24] Miranda Christ and Sam Gunn. Pseudorandom Error-Correcting Codes. 2024. Crypto '24

[K+24] John Kirchenbauer et al. A Watermark for Large Language Models. 2024. PMLR2024