BoundMark: Optimizing Multi-bit LLM Watermark Detectability under Controlled Text Quality Degradation

Chenxi Gu, Si Chen, John Grundy, Xiaoning Du
Monash University

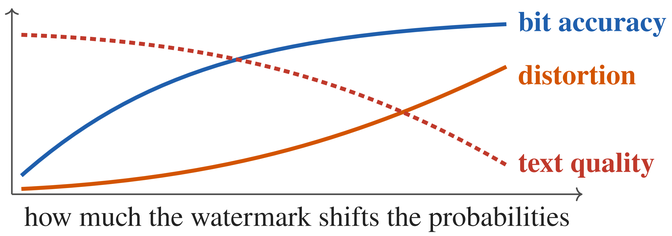
Background

Task: multi-bit watermarking

Embed a **message** (e.g. a user ID) into LLM text: each token carries k bits, and the detector takes a **majority vote** per bit. Success is measured by **bit accuracy (BA)**.

Problem: a trade-off

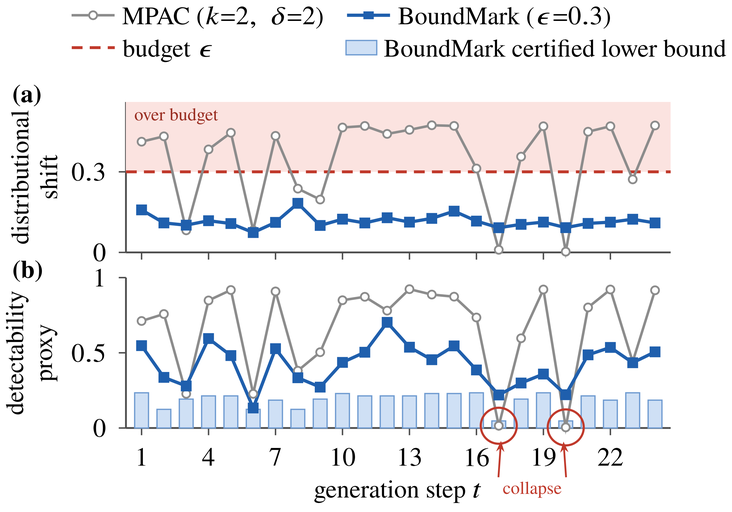
A bigger shift of the next-token distribution makes the message **easier to decode**, but **distorts the distribution** and lowers quality.



Motivation & Challenges

Limitations of existing methods

- **No distortion constraint** (a)
- **No principled way to maximize detectability** (b)



Two challenges

✓ **BA cannot be optimized directly**: only known ***after generation***.

✓ **Distortion constraint**: every step must ***follow the KL budget***.

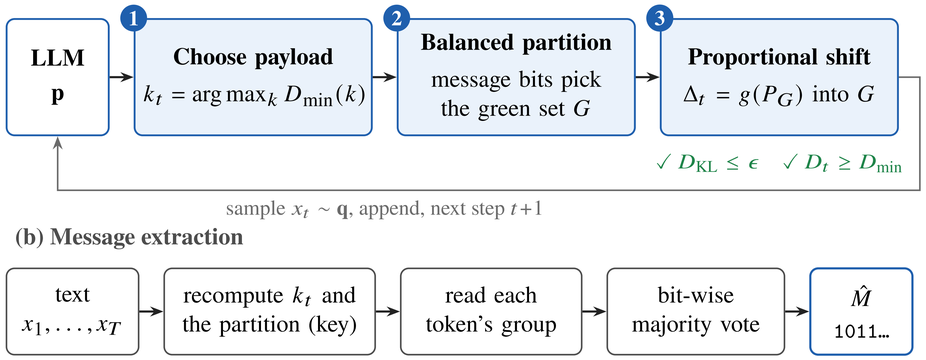
TL;DR

Existing multi-bit watermarks neither **control the distributional distortion** nor **optimize detectability**. **BoundMark** maximizes a ***detectability proxy*** $D = k \cdot \Delta$ within a ***KL budget*** ϵ at every step, with two guarantees:

✓ $D_{\mathrm{KL}}(\mathbf{q}\|\mathbf{p}) \le \epsilon$ distortion within budget · ✓ $D_t \ge D_{\min}(k_t)$ detectability above its lower bound

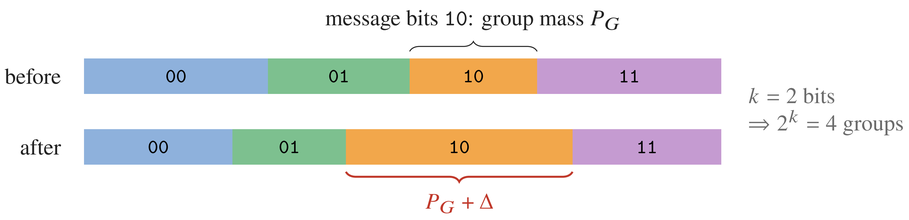
Our Method: BoundMark

Framework



Detectability proxy: $D = k \cdot \Delta$

k: bits carried by the token (payload); Δ: probability moved into the green set G.
G (**green set**): tokens the message bits point to; P_G: its original probability.
m: message length (bits); T: number of generated tokens.



$$D = k \cdot \Delta, \qquad \mathrm{BA} \approx \Phi\left(\Delta\sqrt{Tk/m}\right)$$

Computable at step t, and increasing in both k and Δ.

Nested Optimization

$$\max_{k \ge 0} \; k \cdot \underbrace{\max\{\Delta : \; D_{\mathrm{KL}}(\mathbf{q}\|\mathbf{p}) \le \epsilon\}}_{\text{inner: largest } \Delta \text{ for a given } k}$$

outer: choose k

Inner: the largest Δ for a given k

✓ **Step 1: proportional shift** (Theorem 1, minimal KL)

$$\delta_i = p_i \frac{\Delta}{P_G} \;\; (i \in G), \qquad \delta_j = -p_j \frac{\Delta}{1-P_G} \;\; (j \in R)$$

✓ **Step 2: maximal shift** (Theorem 2)

$$\Delta_{\max} = g(P_G) = \min\left\{\sqrt{\epsilon P_G(1-P_G)},\; 1-P_G\right\}$$

Outer: choose k

✓ **Step 3: maximize a lower bound** (exact P_G is **intractable**)

$$k_t = \arg\max_{0 \le k \le k_{\max}} D_{\min}(k), \qquad \Delta_t = g(P_G)$$

Code & Contact

Code (GitHub) **LinkedIn**