Can Watermarked LLMs be Identified by Users via Crafted Prompts?

Aiwei Liu, Sheng Guan, Yiming Liu, Leyi Pan, Yifei Zhang, Liancheng Fang, Lijie Wen, Philip S. Yu, Xuming Hu

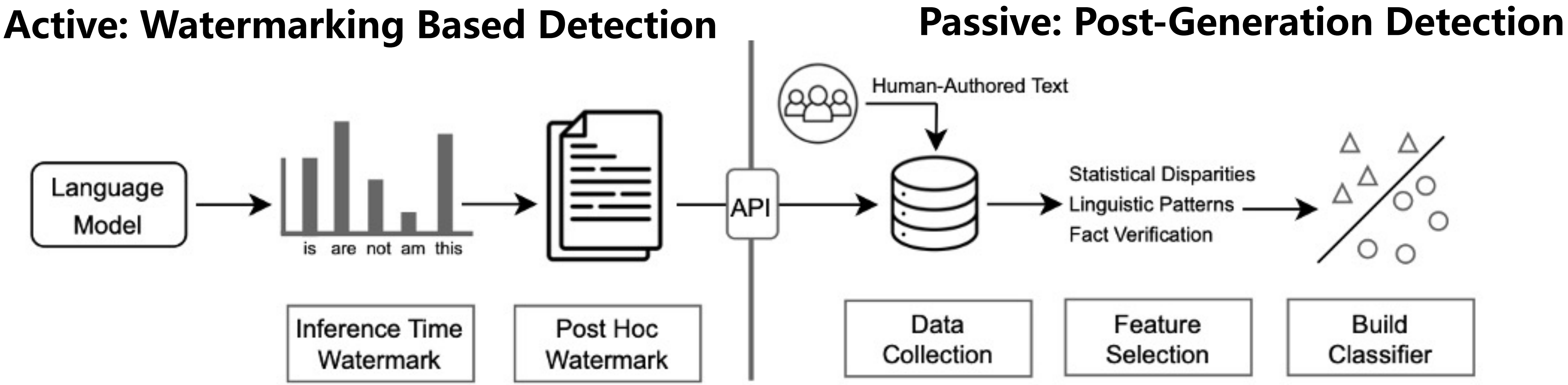
ICLR 2025 Spotlight

2025.04.25

Background Watermark For LLMs

Large language models can rapidly generate text that may cause harmful effects.

The text generated by LLMs needs to be detected and tracked !

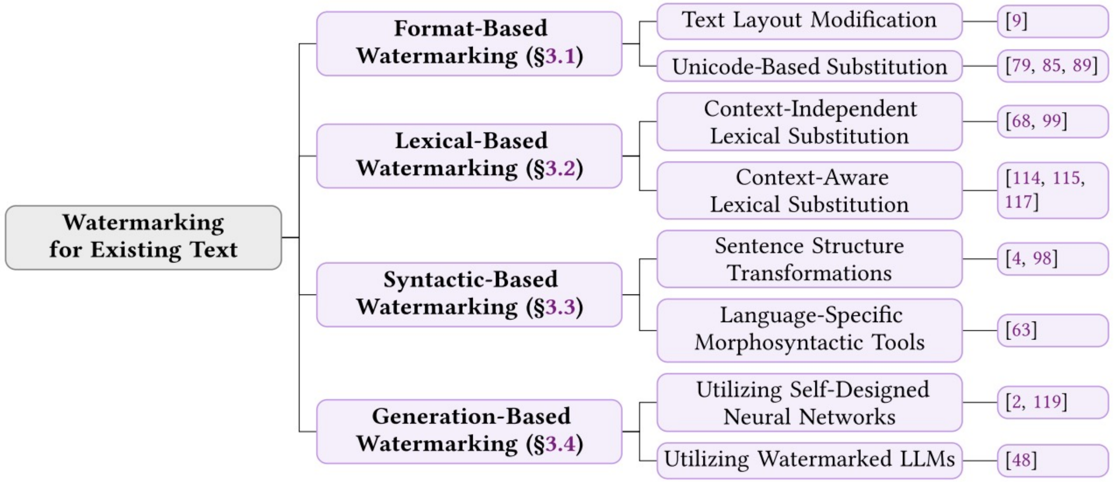


Watermarking for large language models is a more reliable method for detecting and tracking AI-generated text.

History of Text Watermarking

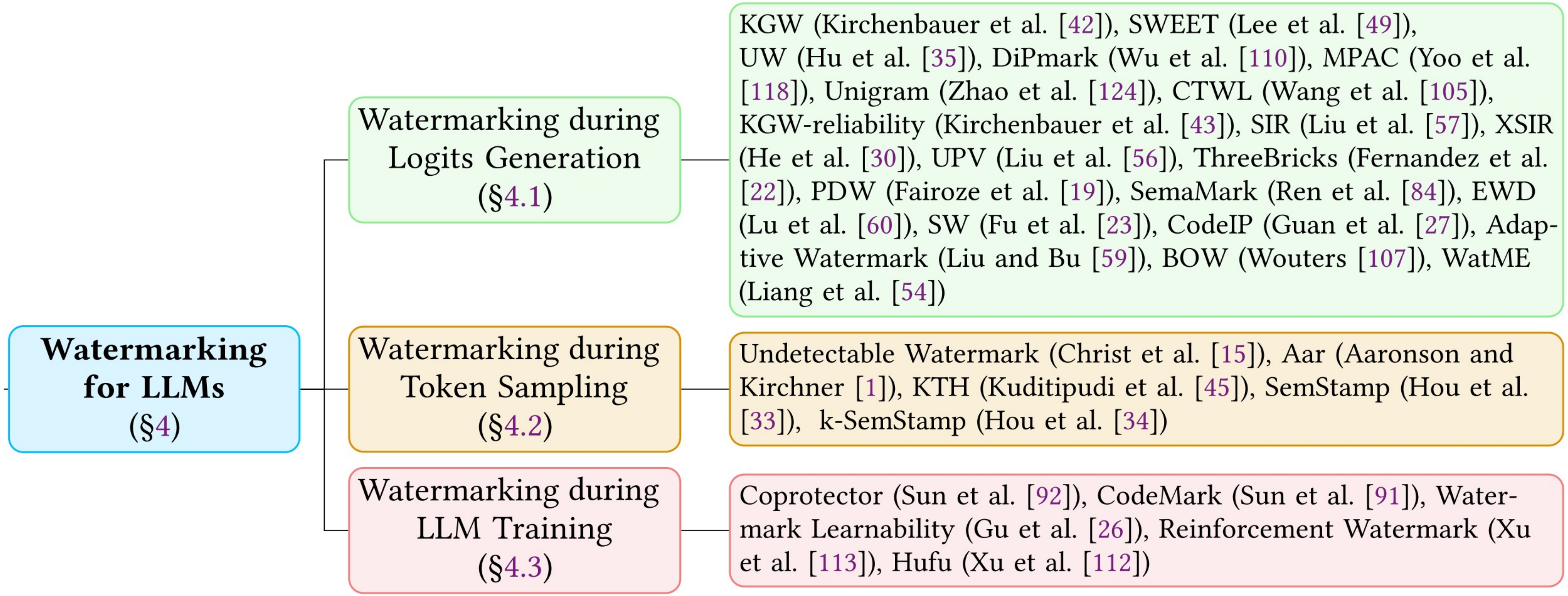
- Ancient Greece: Steganography

- 1950s: Embedding code to music (Hembrooke, 1954)
- 1990s to 2000s: Digital Watermarks (e.g., Ingemar J. Cox, Matt Miller, etc..)
- Rule-based parsed syntactic tree (Atallah et al., 2001)
- Rule-based semantic structure of text (Atallah et al., 2000; Topkara et al., 2006)
- Neural steganography with DL models (Fang et al., 2017; Ziegler et al., 2019)



[1] Aiwei Liu, et al. "A survey of Text Watermarking in the era of Large Language Models." ACM Computing Survey

2022+: Recent Renaissance due to the rise of Generative AI



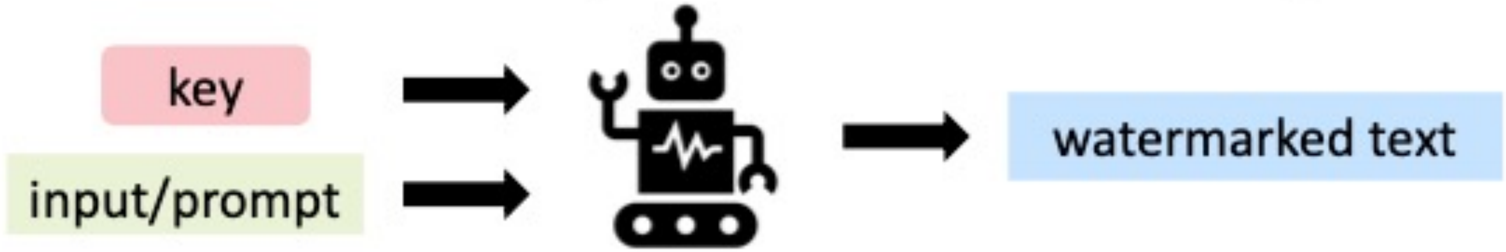
Traditional Method: Given text, change text to add watermark.

Modern LLM Text Watermark: We also have access to the original generative process.

[1] Aiwei Liu, et al. "A survey of Text Watermarking in the era of Large Language Models." ACM Computing Survey

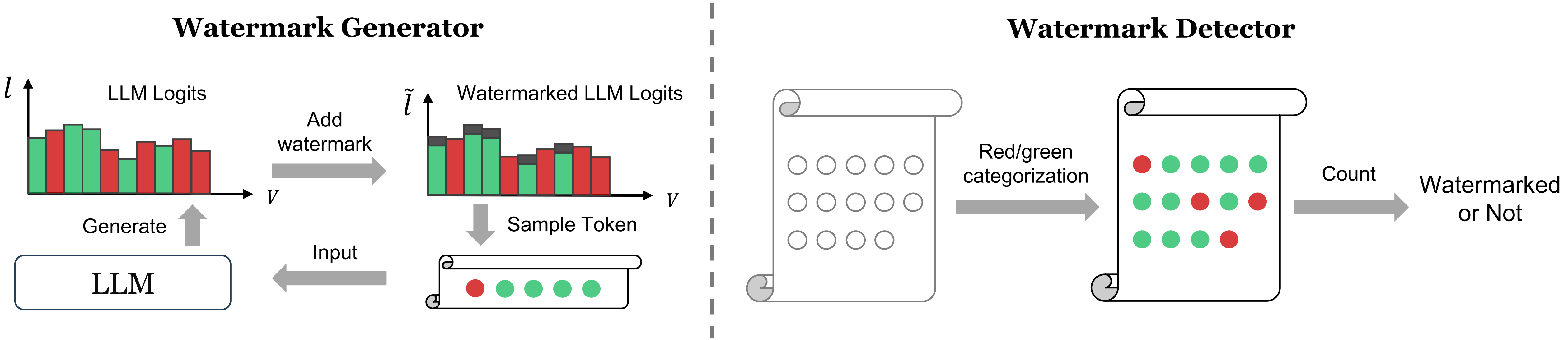
What is an LLM Text Watermarking?

- Watermark($\mathcal{M}$): (possibly randomized procedure) that outputs a new model $\hat{\mathcal{M}}$, and detection key $\boldsymbol{k}$



- Detect($\boldsymbol{k}, \boldsymbol{y}$): takes input detection key $\boldsymbol{k}$ and sequence $\boldsymbol{y}$, then outputs 1 (indicating it was AI-generated) or 0 (indicating it was human-generated)

Example Method: KGW(Red-Green) Watermark



The KGW (Kirchenbauer et al. 2023)[1] watermarking algorithm divides the vocabulary into red and green token lists and embeds watermarks by slightly increasing the probability of green list tokens.

[1] Kirchenbauer, John, et al. "A watermark for large language models." ICML 2023

Paradigm 1: N-gram watermarking

KGW is an **N-gram watermark**.

The green list G at each step is determined by previous $(\mathrm{N}-1)$ tokens:

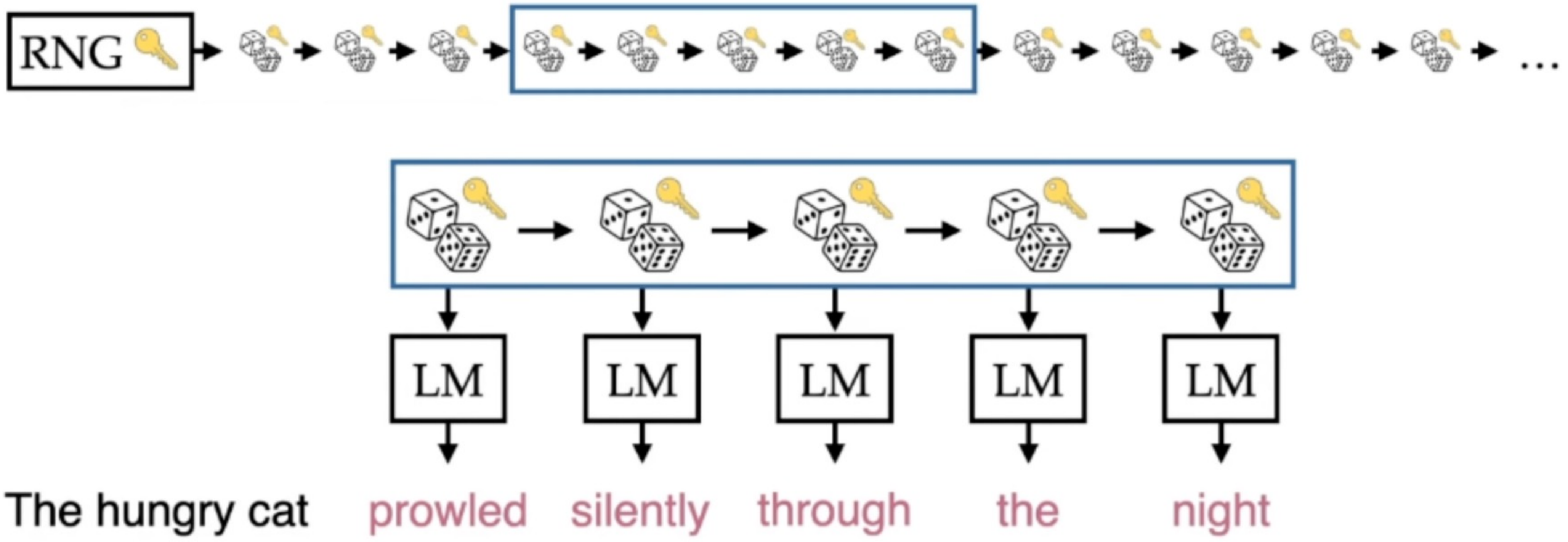
$$G(x_{1:N-1}) \subseteq V$$

Two implementations:

- KGW watermark (Kirchenbauer et al.)[1]
  : N = 2, green list determined by previous one token
- Unigram (Zhao et al.)[2]: N = 1, a constant green list

[1] Kirchenbauer, John, et al. "A watermark for large language models." ICML 2023
[2] Zhao, Xuandong, et al. "Provable robust watermarking for ai-generated text." ICLR 2024

Paradigm 2: Fixed Key list based watermarking



Unlike previous N-gram generated watermark keys, Fixed Key list based watermarking provides a predefined watermark key list, randomly selecting a starting position during generation and proceeding sequentially.

Problem: How could we know if a LLM service is watermarked?

deepseek

OpenAI

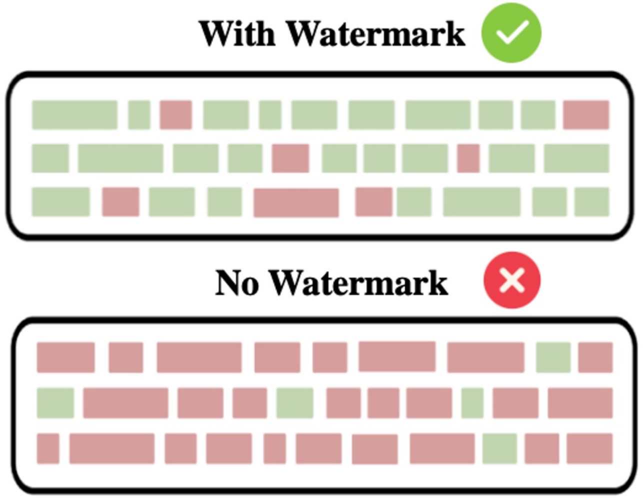
GPT - 4

Gemini

Claude BY ANTHROP\C

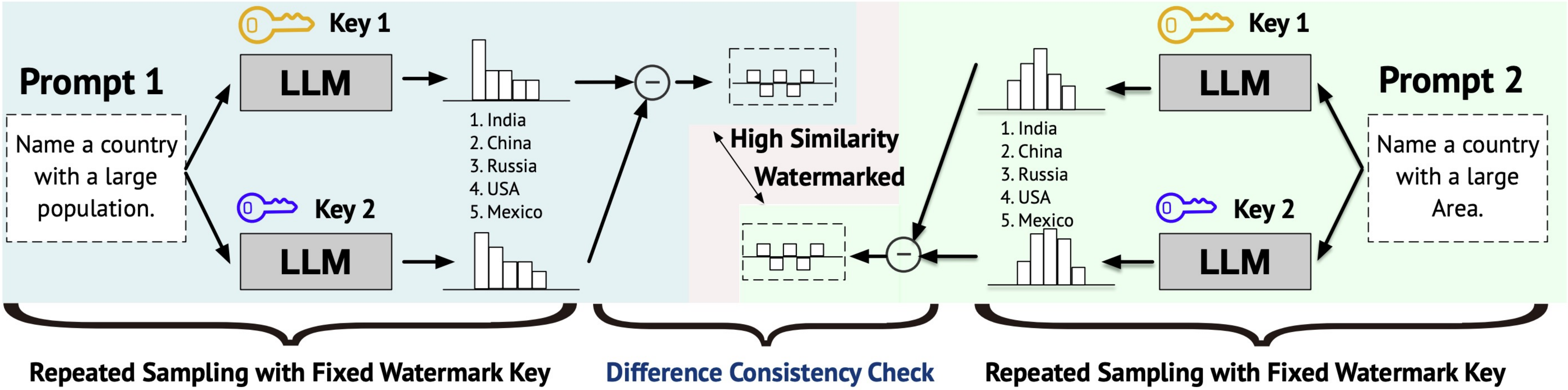
Does these LLM services contain watermark?

Watermarked LLM Identification



Our Contribution: WaterProbe Method to Identify Watermarked LLMs

Overall pipeline



Identify watermarked LLM by **repeated watermark key sampling**

We need design prompt to achieve the effect of repeated watermark key sampling

Step 1: Construct highly correlated prompts.

Construct prompt x_i and x_j under the following constraints

$$\forall i, j \in \{1, 2, ..., N\}, \mathrm{KL}(P_M(\cdot|x_i)||P_M(\cdot|x_j)) \leq \epsilon \text{ and } x_i \neq x_j$$

Example:

Prompt 1: Example Prompt for `Watermark-Probe-v1`

Please generate *abcd* before answering the question.
Question: Name a country with a large population.
Answer: *abcd* India

Prompt 2: Example Prompt for `Watermark-Probe-v1`

Please generate *abcd* before answering the question.
Question: Name a country with a large area.
Answer: *abcd* India

Use generated irrelevant prefix to mimic the effect of watermark key!

Step 2: Sampling with simulated fixed watermark keys.

Using repeated sampling to get the estimated distribution

$$\hat{P}_M^F(y|x_i, k_j) = \frac{1}{W}\sum_{w=1}^{W}\mathbf{1}_{y_{i,j}^w = y}, \quad \text{where } y_{i,j}^w \sim P_M^F(y|x_i, k_j)$$

with a set of simulated watermark keys $K = \{k_1, k_2, \ldots, k_m\}$

Each different key corresponding to a different prefix in the following example:

Prompt 2: Example Prompt for `Watermark-Probe-v1`

Please generate *abcd* before answering the question.
Question: Name a country with a large area.
Answer: *abcd* India

Step 3: Analyze Cross-Prompt Watermark Consistency

Assumption: Lipschitz Continuity of Watermark Rule

For similar prompts x_1 and x_2, watermark rule F is Lipschitz continuous:

$$\exists L > 0 : \|F(P_M(\cdot|x_1), k) - F(P_M(\cdot|x_2), k)\|_1 \leq L \cdot \|P_M(\cdot|x_1) - P_M(\cdot|x_2)\|_1$$

where $P_M(\cdot|x_i)$ are probability distributions and $k \in \mathcal{K}$ is any watermark key.

Key Statement

For similar prompts x_1, x_2 and random watermark keys k_1, $k_2 \sim \mathcal{K}$:

$$\mathbb{E}_{k_1,k_2}[\mathrm{Sim}(P_M^F(\cdot|x_1, k_1) - P_M^F(\cdot|x_1, k_2), P_M^F(\cdot|x_2, k_1) - P_M^F(\cdot|x_2, k_2))] \geq \rho$$

where:

- P_M^F is the watermarked distribution
- $\mathrm{Sim}(\cdot, \cdot)$ is a similarity measure
- ρ is a constant significantly greater than 0

Waterprob v2: Identify all watermark Paradigm

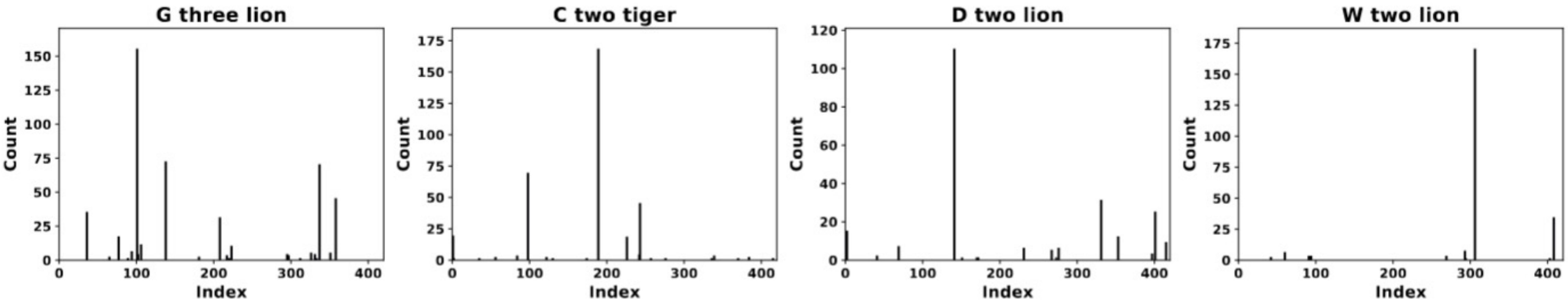
Previous introduced prompt example could only identify the N-gram based watermark paradigm, the following prompt could help identify all paradigm.

Prompt 2: Example Prompt for `Watermark-Probe-v2`

Please generate a sentence that satisfies the following conditions: The first word is randomly sampled from *A-Z*. The second word is randomly sampled from *zero to nine*. The third word is randomly sampled from *cat, dog, tiger and lion*. Then answer the question: Name a country with a large population.

Answer: *A one cat* China

Start key distribution under different prefix in the fixed watermark key list paradigm



Experiment Result On Opensource LLMs

LLM	N-Gram							Fixed-Key-List	
	Non	KGW	Aar	KGW-Min	KGW-Skip	DiPmark	γ-Reweight	EXP-Edit	ITS-Edit
Water-Probe-v1 (w. prompt 2)									
Qwen2.5-1.5B	0.02 ± 0.02	0.37 ± 0.02	0.88 ± 0.06	0.37 ± 0.02	0.39 ± 0.01	0.55 ± 0.01	0.55± 0.01	0.01± 0.02	0.00± 0.04
OPT-2.7B	0.05 ± 0.01	0.47 ± 0.01	0.91 ± 0.01	0.42 ± 0.02	0.45 ± 0.01	0.60 ± 0.01	0.61± 0.01	0.08 ± 0.02	0.09± 0.01
Llama-3.2-3B	0.04 ± 0.02	0.53 ± 0.01	0.90 ± 0.01	0.48 ± 0.00	0.49 ± 0.01	0.61 ± 0.01	0.61± 0.01	0.03 ± 0.01	0.04± 0.01
Qwen2.5-3B	0.03 ± 0.01	0.33± 0.02	0.75± 0.05	0.33± 0.02	0.38 ± 0.00	0.51± 0.01	0.53± 0.01	0.03± 0.01	0.06± 0.02
Llama2-7B	0.02 ± 0.01	0.42 ± 0.01	0.87 ± 0.01	0.31 ± 0.01	0.42 ± 0.01	0.56 ± 0.01	0.56± 0.04	0.03 ± 0.02	0.02± 0.00
Mixtral-7B	0.01 ± 0.02	0.41 ± 0.01	0.85 ± 0.02	0.37 ± 0.01	0.41 ± 0.02	0.57 ± 0.01	0.58± 0.03	0.00± 0.00	0.02± 0.02
Qwen2.5-7B	0.07 ± 0.04	0.41 ± 0.02	0.82 ± 0.02	0.34 ± 0.03	0.38 ± 0.02	0.43 ± 0.03	0.43± 0.02	0.06± 0.01	0.04± 0.02
Llama-3.1-8B	0.01 ± 0.02	0.41 ± 0.02	0.85 ± 0.02	0.41 ± 0.01	0.39 ± 0.01	0.57 ± 0.02	0.58± 0.00	0.02± 0.02	0.00± 0.01
Llama2-13B	0.01 ± 0.03	0.41 ± 0.01	0.86 ± 0.01	0.31 ± 0.02	0.40 ± 0.02	0.58 ± 0.02	0.60± 0.01	0.02 ± 0.01	0.02± 0.03
Average	0.029	0.418	**0.854**	0.371	0.412	0.553	0.505	0.031	0.032
Water-Probe-v2 (w. prompt 3)									
Qwen2.5-1.5B	0.02± 0.02	0.30± 0.01	0.83± 0.01	0.29± 0.01	0.27± 0.02	0.49± 0.02	0.52± 0.03	0.39± 0.03	0.60 ± 0.00
OPT-2.7B	0.04 ± 0.03	0.29 ± 0.02	0.88 ± 0.01	0.23 ± 0.01	0.19 ± 0.02	0.42 ± 0.01	0.43± 0.03	0.43 ± 0.01	0.62 ± 0.00
Llama-3.2-3B	0.00 ± 0.01	0.31 ± 0.01	0.89 ± 0.01	0.33 ± 0.00	0.24 ± 0.01	0.51 ± 0.01	0.54± 0.01	0.52 ± 0.01	0.84 ± 0.00
Qwen2.5-3B	0.03± 0.02	0.35± 0.04	0.78± 0.01	0.29± 0.02	0.28± 0.01	0.45± 0.02	0.45± 0.02	0.39± 0.02	0.71 ± 0.00
Llama2-7B	0.04 ± 0.02	0.34 ± 0.01	0.82 ± 0.02	0.33 ± 0.01	0.28 ± 0.01	0.50 ± 0.01	0.51± 0.02	0.48 ± 0.01	0.81 ± 0.00
Mixtral-7B	0.09 ± 0.01	0.34 ± 0.04	0.83 ± 0.01	0.29 ± 0.02	0.24 ± 0.01	0.51 ± 0.01	0.53± 0.00	0.42± 0.02	0.81 ± 0.00
Qwen2.5-7B	-0.01 ± 0.04	0.26 ± 0.02	0.70 ± 0.00	0.28 ± 0.02	0.23 ± 0.01	0.32 ± 0.03	0.35± 0.02	0.32± 0.02	0.73 ± 0.00
Llama-3.1-8B	0.01 ± 0.00	0.31 ± 0.01	0.77 ± 0.01	0.29 ± 0.02	0.26 ± 0.00	0.50 ± 0.01	0.51 ± 0.01	0.43± 0.01	0.71 ± 0.00
Llama2-13B	0.01 ± 0.02	0.35 ± 0.01	0.82 ± 0.02	0.26 ± 0.02	0.26 ± 0.01	0.50 ± 0.01	0.53± 0.01	0.44 ± 0.02	0.73 ± 0.00
Average	0.026	0.317	**0.813**	0.288	0.250	0.467	0.486	0.424	0.729

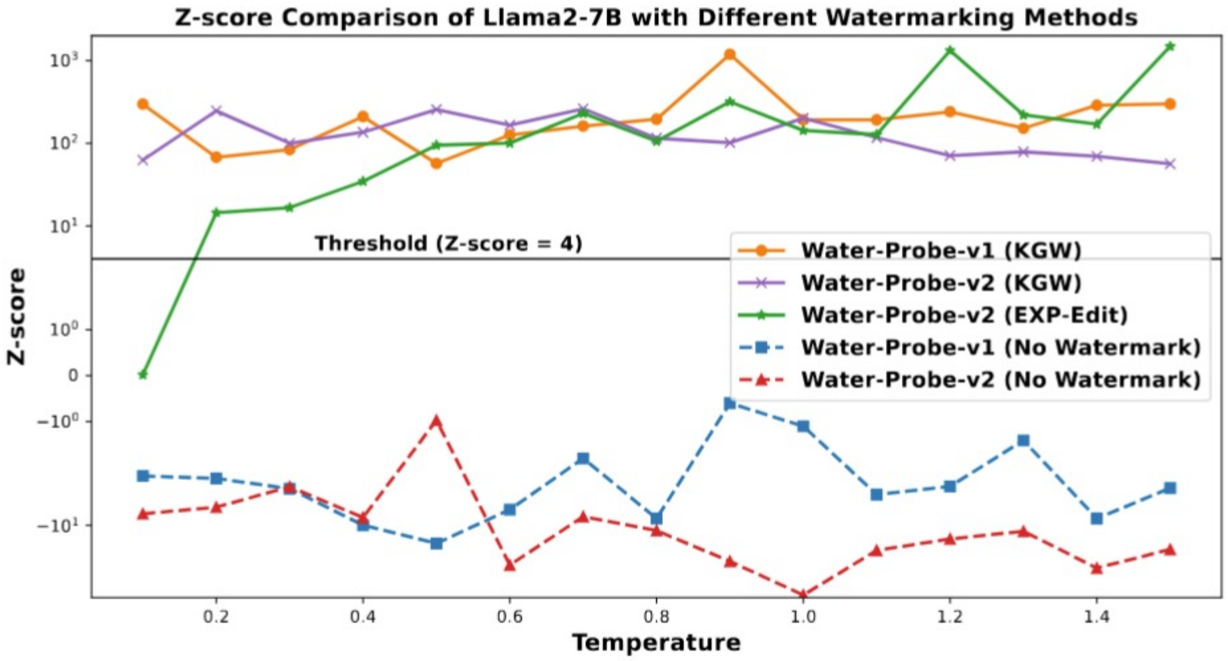
Water-Probe-V2 Method Could identify all watermark method for different kind of LLMs.

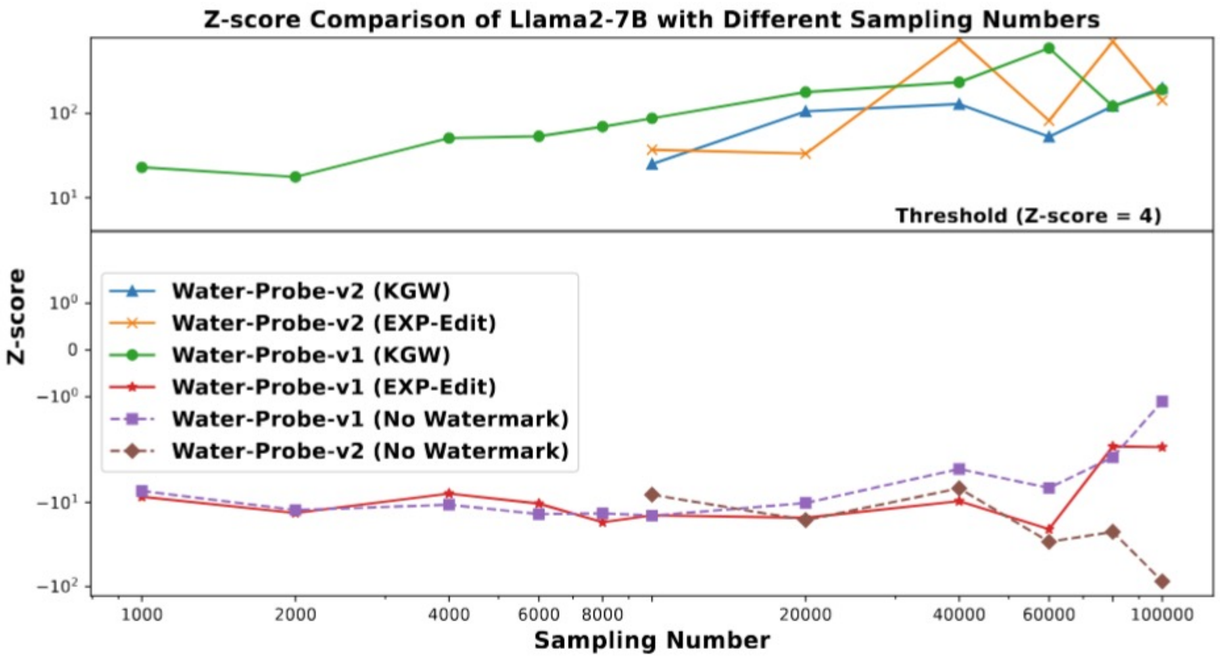
Experiment Result On Commercial LLMs

Model	Similarity	Std Dev	Z-score	Watermarked?
GPT-4o-mini	-0.005	0.018	-5.984	No
GPT-4o	0.017	0.020	-4.211	No
GPT-3.5-turbo	0.028	0.030	-2.362	No
Gemini-1.5-flash	0.027	0.049	-1.474	No
Gemini-1.5-pro	0.018	0.038	-2.135	No

No watermark identified in current commercial LLMs

Experiment Result: Further Analysis





Can perform well at different temperature settings.

Only require 1000 samples to identify watermarked LLM.

Prevent Watermarked LLM from Being Detected: Waterbag strategy

Key Components

- Uses master keys $K = \{K_1, ..., K_n\}$ and inversions $\overline{K} = \{\overline{K_1}, ..., \overline{K_n}\}$
- For each generation, randomly selects K_j or $\overline{K_j}$

Modified Distribution

$$P_M^{WB}(y_i|x, y_{1:i-1}, K, \overline{K}) = F(P_M(y_i|x, y_{1:i-1}), k_i)$$

where $k_i = f(K_j^*, y_{i-n:i-1})$, $K_j^* \sim \text{Uniform}(K \cup \overline{K})$

Inversion Property

$$\frac{1}{2}(F(P_M(\cdot), f(K_j, \cdot)) + F(P_M(\cdot), f(\overline{K_j}, \cdot))) = P_M(\cdot)$$

Ensures average effect of key pairs equals original distribution

Experiment result for waterbag strategy

	None	KGW w. Water-Bag				Exp-Edit(Key-len)		
		$\lvert K \cup \overline{K} \rvert = 1$	$\lvert K \cup \overline{K} \rvert = 2$	$\lvert K \cup \overline{K} \rvert = 4$	$\lvert K \cup \overline{K} \rvert = 8$	$\lvert K \rvert = 420$	$\lvert K \rvert = 1024$	$\lvert K \rvert = 2048$
Watermarked LLM Indentification								
Water-Probe-v1(n=3)	0.02 ± 0.01	0.42 ± 0.01	0.05 ± 0.01	0.02 ± 0.01	0.03 ± 0.02	0.03 ± 0.05	0.02 ± 0.01	0.02 ± 0.02
Water-Probe-v2(n=3)	0.04 ± 0.01	0.34 ± 0.01	0.34 ± 0.01	0.25 ± 0.01	0.16 ± 0.02	0.48 ± 0.01	0.33±0.01	0.23±0.02
Water-Probe-v2(n=5)	0.06 ± 0.06	0.32 ± 0.01	0.18 ± 0.01	0.12 ± 0.02	0.07 ± 0.01	0.64 ± 0.00	0.54 ± 0.01	0.44 ± 0.00
Watermarked Text Detection								
Detection-F1-score	-	1.0	1.0	1.0	1.0	1.0	0.975	1.0
PPL	8.15	11.93	11.85	12.17	12.50	16.63	17.28	19.06
Robustness (GPT3.5)	-	0.843	0.849	0.748	0.696	0.848	0.854	0.745
Detection-time (s)	-	0.045	0.078	0.156	0.31	37.87	108.5	194.21

After implementing the waterbag strategy, watermarked LLMs become difficult to detect, while maintaining their inherent detectability, robustness, and other properties.

Thank You!

Thank You!