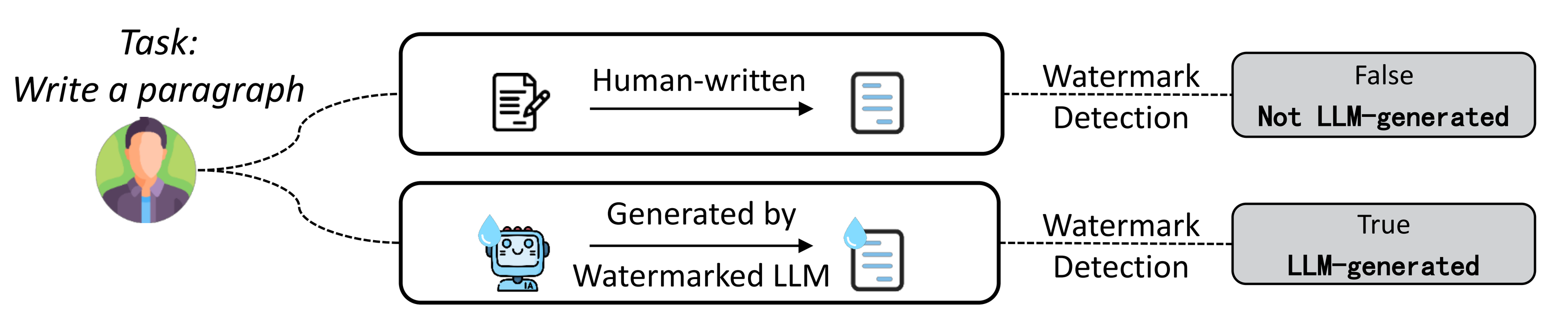
Can Watermarked LLMs be Identified by Users via Crafted Prompts?

Aiwei Liu[1], Sheng Guan[2], Yiming Liu[1], Leyi Pan[1], Yifei Zhang[3], Liancheng Fang[4], Lijie Wen[1], Philip S. Yu[4], Xuming Hu[5]

[1]Tsinghua University, [2]Beijing University of Posts and Telecommunications, [3]The Chinese University of Hong Kong, [4]University of Illinois at Chicago, [5]HKUST (Guangzhou)

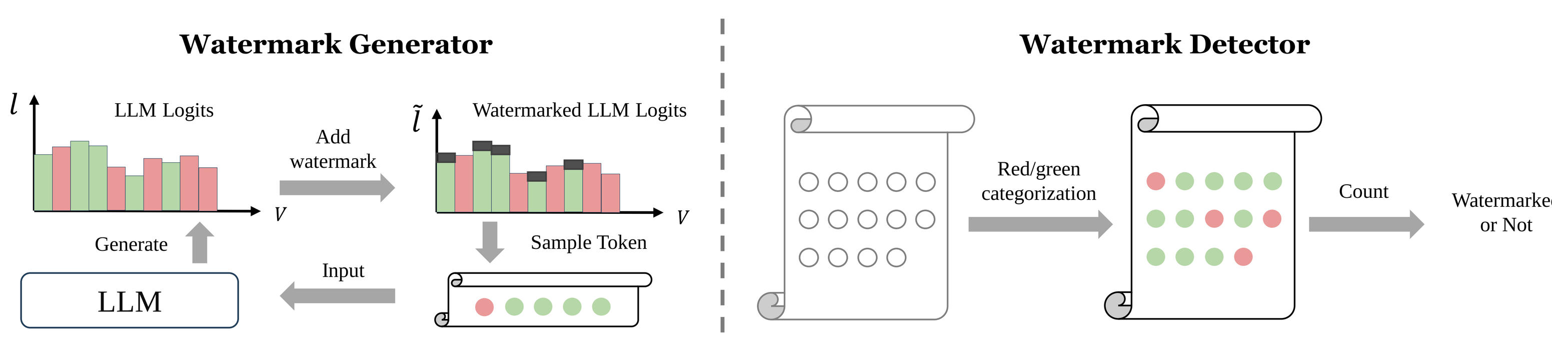
What is LLM Watermarking?

Imperceptible features are embedded in text generated by **large language models (LLMs)** to identify LLM-generated content.



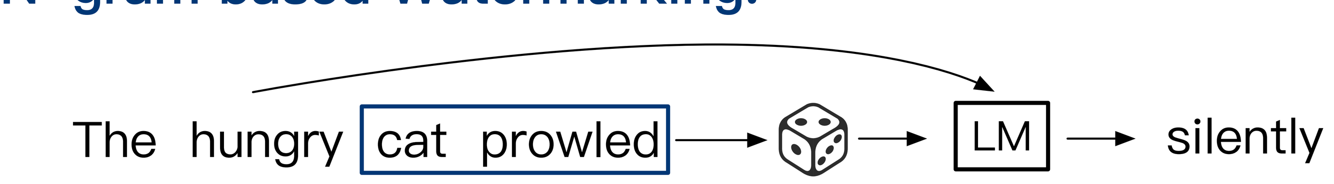
Text Generation and Detection Process

Example: KGW watermarking algorithm, which split the vocabulary into red and green list, and add the probability of the green list tokens.

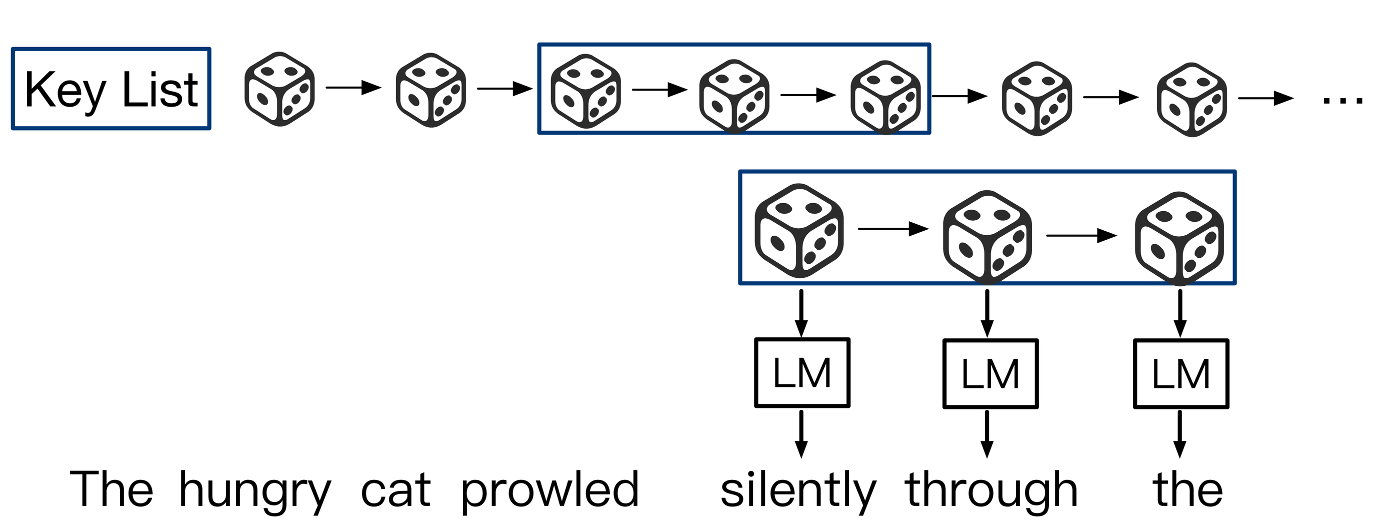


Two type of watermarking: N-gram based and fixed key list based. (Depends on how to get the watermark key to split the vocabulary)

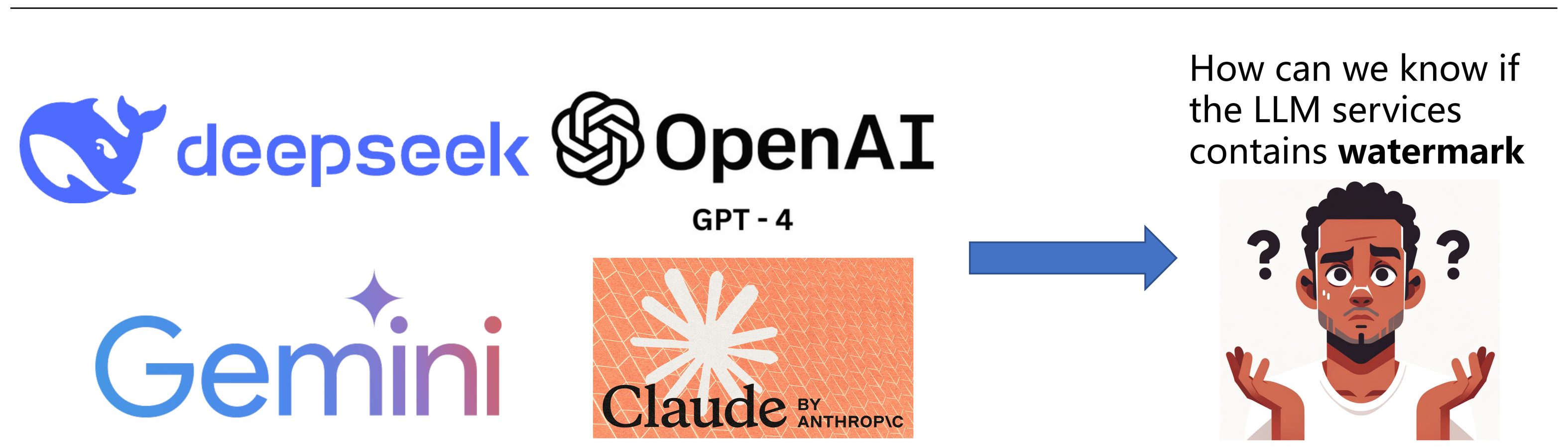
N–gram based Watermarking:



Fixed Key–list based Watermarking:



Problem: How to Identify Watermarked LLM?



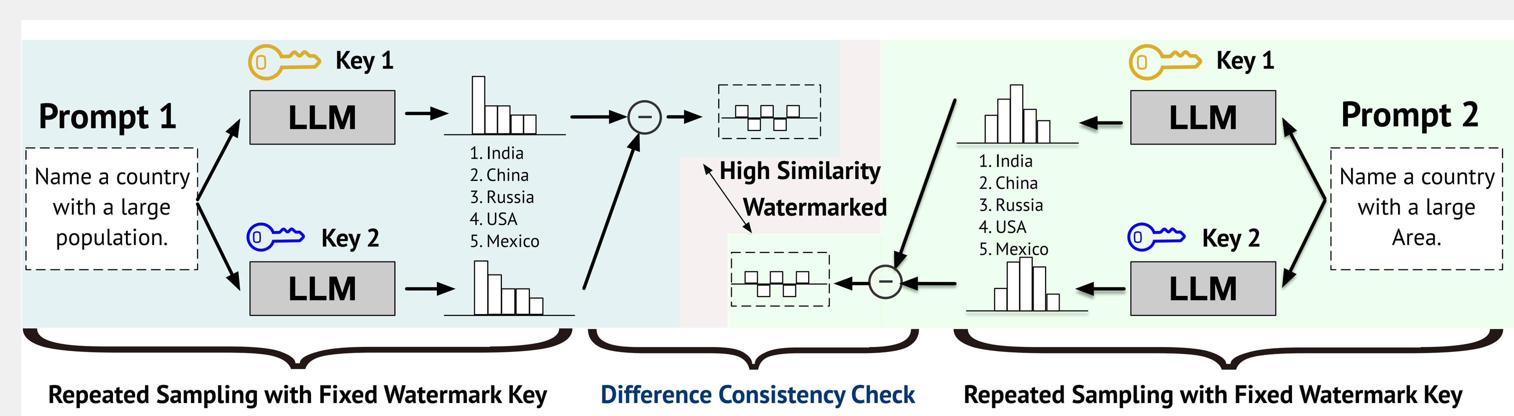
Water-Probe Method to Identify Watermarked LLMs

Key Idea:

Detect watermarks by analyzing distribution differences under repeated key sampling

Method:

Construct highly correlated prompts with similar output distributions
Sample outputs with simulated watermark keys
Analyze cross-prompt watermark consistency using rank transformation
Use z-test to determine if LLM is watermarked



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-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

Details of Watermark Consistency Check

Step 1: Calculate average similarity across prompts and keys:

$$\bar{S} = \frac{1}{N}\sum_{P_i \neq P_j}\sum_{k_m \neq k_n} \mathrm{Sim}(\Delta_R(P_i, k_m, k_n), \Delta_R(P_j, k_m, k_n)) \tag{1}$$

where $\Delta_R(P, k_m, k_n) = R(\hat{P}^F_M(\cdot|P, k_m)) - R(\hat{P}^F_M(\cdot|P, k_n))$ measures rank difference between key pairs.

Step 2: Perform statistical test:

$$z = (\bar{S} - \mu)/\sigma \tag{2}$$

If $z > z_\alpha$, conclude watermarked (high consistency indicates watermark).

Water-Bag: Improve the Imperceptibility of Watermark

Key Idea: Enhance watermark imperceptibility by using **multiple master keys** with **inversion mechanism**. The core principle is to make it challenging to construct **repeated sampling scenarios** using different keys.

Method: For each generation, randomly select a key K_j or its inversion $\overline{K_j}$ from a **key-bag**:

$$P_M^{WB}(y_i|x, y_{1:i-1}) = F(P_M(y_i|x, y_{1:i-1}), k_i), \quad k_i = f(K_j^*, y_{i-n:i-1}) \tag{3}$$

where K_j^* is randomly sampled from the combined set of **original and inverted keys**. The inverted key $\overline{K_j}$ ensures that the **average effect** of keys equals the original distribution.

Water-Probe: Experiment Results

Experiment on Open-Source LLMs with different watermarks.

LLM	N-Gram					Fixed-Key-List	
	Non	KGW	Aar	DiPmark	γ-Reweight	EXP-Edit	ITS-Edit
`Water-Probe-v1` (w. prompt 1)							
Qwen2.5-1.5B	0.02 ± 0.02	0.37 ± 0.02	0.88 ± 0.06	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	0.60 ± 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	0.61 ± 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	0.51 ± 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	0.56 ± 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	0.57 ± 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	0.43 ± 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	0.57 ± 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	0.58 ± 0.02	0.60 ± 0.01	0.02 ± 0.01	0.02 ± 0.03
Average	0.029	0.418	**0.854**	0.553	0.505	0.031	0.032
`Water-Probe-v2` (w. prompt 1)							
Qwen2.5-1.5B	0.02 ± 0.02	0.30 ± 0.01	0.83 ± 0.01	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	0.42 ± 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	0.51 ± 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	0.45 ± 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	0.50 ± 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	0.51 ± 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	0.32 ± 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	0.50 ± 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	0.50 ± 0.01	0.53 ± 0.01	0.44 ± 0.02	0.73 ± 0.00
Average	0.026	0.317	**0.813**	0.467	0.486	0.424	0.729

Experiment on closed-source 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

liuaw20@mails.tsinghua.edu.cn