WatME: Towards Lossless Watermarking Through Lexical Redundancy

香港中文大學
The Chinese University of Hong Kong

Liang Chen, Yayao Bian, Yang Deng, Deng Cai, Shuaiyi Li, Peilin Zhao, Kam-Fai Wong

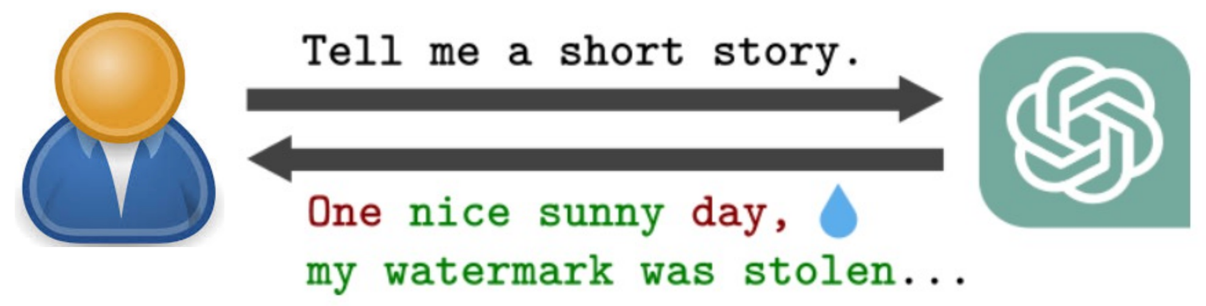
The Chinese University of Hong Kong, National University of Singapore, Tencent AI Lab

lchen@se.cuhk.edu.hk

NUS
National University of Singapore

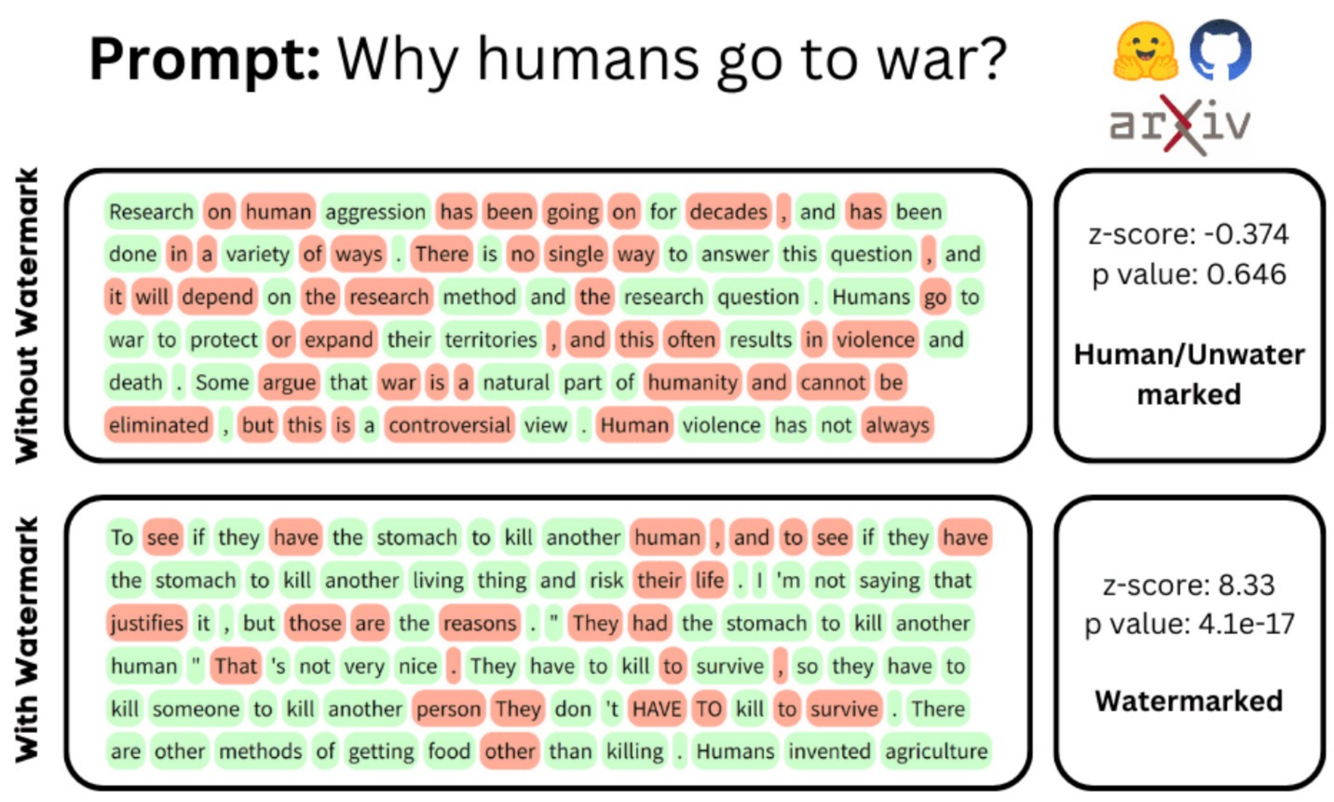
❖ Research Background

➢ Watermarking is the most effective method for detecting machine-generated text.



➢ How does watermarking work?

- **Embed watermark:** At every step, subtly bias LLM logits by partitioning vocabulary into green and red sets. Increase sampling probability for green tokens.
- **Detect watermark:** Observe a high number of green tokens, indicating the presence of a watermark.



➢ Challenges of text watermarking

- **Severely impairs response quality:** Relies on arbitrary vocabulary partitioning during decoding, potentially leaving no suitable words available.
- **Discrete nature of text data:** Unlike images with redundant pixels for watermarking, text is discrete and concise, offering almost no redundant space.

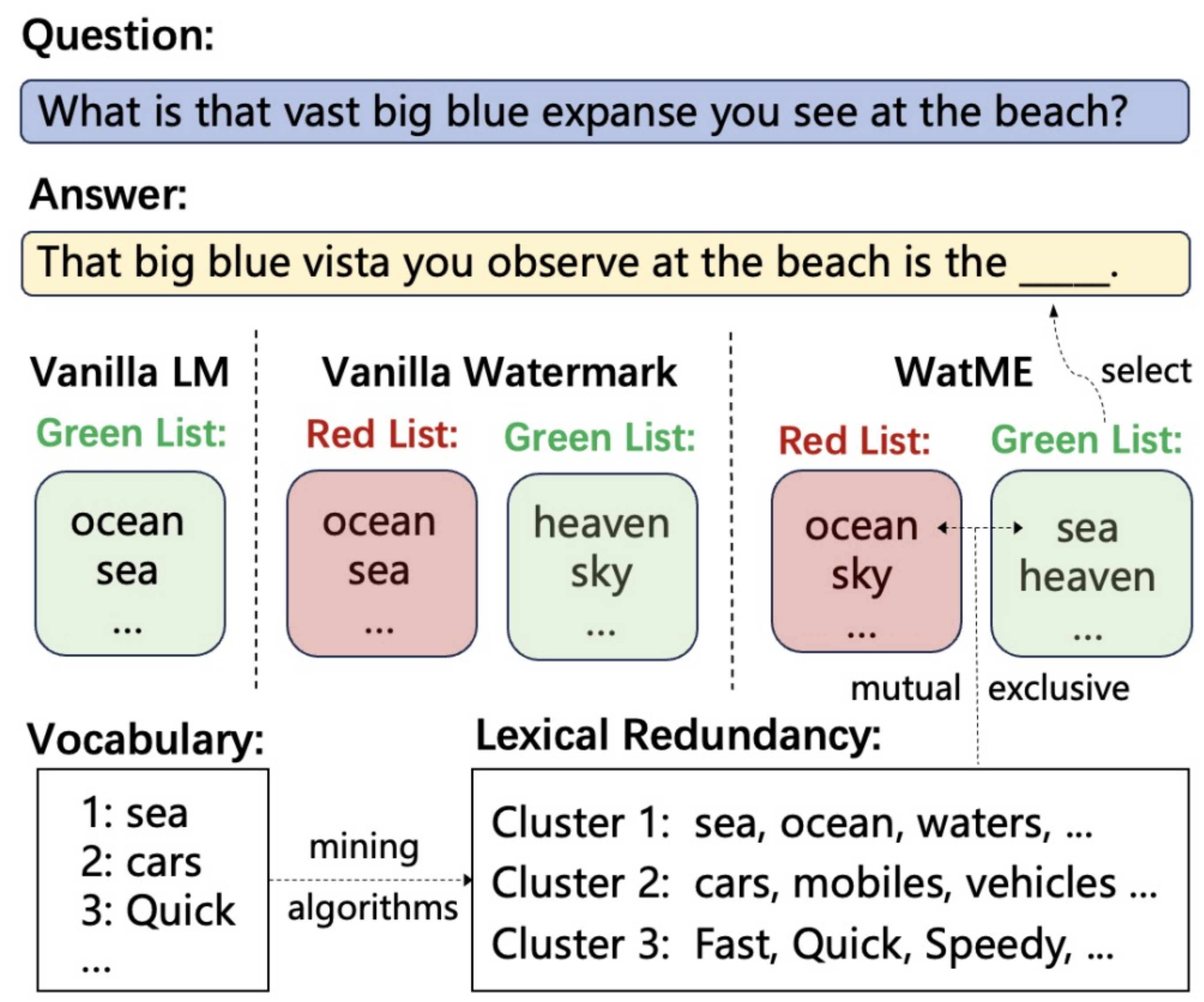
❖ Motivation & Method

➢ Motivation

- Inspired by image watermarking, we propose identifying redundancy within data to enable lossless watermarking.

➢ A related concept: lexical redundancy

- LLM vocabulary contains many tokens with similar semantic and syntactic functions. Some can be disabled while others substitute.
- This redundancy creates space to embed watermarks.



➢ Use lexical redundancy in Watermarking

- **Explore:** We constructed structured redundancy clusters using LLM-based and dictionary-based methods.
- **Exploit:** When embedding watermarks, we first partition the redundancy clusters, then divide the remaining vocabulary. Maximizing the partitioning of redundant elements minimizes the impact of watermarking.

❖ Theoretical Justification

➢ We demonstrate the advantages of our method through two theories:

- **Theory 1 (informal):** Our method increases the likelihood of selecting suitable tokens at each decoding step.
- **Theory 2 (informal):** Our method more effectively preserves the language model's expressiveness.

❖ Empirical Validation

➢ Our method beats baselines on 3 tasks.

Model	GSM8K		TruthfulQA				C4	
	Acc.	AUROC	True.	Info.	True.*Info.	AUROC	PPL	AUROC
LLAMA2-7B	11.22	-	95.10	92.78	88.23	-	4.77	-
+ KGW-MARK	$5.61_{-50.0\%}$	0.8886	$57.16_{-39.9\%}$	$84.33_{-9.1\%}$	$48.20_{-45.4\%}$	0.8416	7.00	0.9724
+ GUMBEL-MARK	$7.28_{-35.1\%}$	0.9121	$45.90_{-51.7\%}$	$92.78_{-0.0\%}$	$42.59_{-51.7\%}$	0.4931	39.93	0.9422
+ UNBIASED-MARK	$10.24_{-8.7\%}$	0.5478	$44.06_{-53.7\%}$	$93.76_{+1.1\%}$	$41.43_{-53.0\%}$	0.5051	15.62	0.5451
+ PROVABLE-MARK	$5.16_{-54.01\%}$	0.9052	$64.14_{-32.6\%}$	$91.68_{-1.2\%}$	$58.80_{-33.4\%}$	0.9555	10.21	0.9623
+ WATME$_{dictionary}$	$9.17_{-18.3\%}$	0.8995	$69.28_{-27.2\%}$	$88.25_{-4.9\%}$	$61.14_{-30.7\%}$	0.8848	5.32	0.9804
+ WATME$_{prompting}$	$5.84_{-48.0\%}$	0.9128	$55.83_{-41.3\%}$	$95.10_{+2.5\%}$	$50.39_{-42.9\%}$	0.8659	6.89	0.9724
VICUNA-V1.5-7B	17.51	-	93.88	87.27	81.92	-	10.77	-
+ KGW-MARK	$13.87_{-20.8\%}$	0.7870	$74.05_{-21.1\%}$	$87.52_{+0.3\%}$	$64.81_{-20.1\%}$	0.7417	11.62	0.9679
+ GUMBEL-MARK	$9.02_{-48.5\%}$	0.7077	$68.30_{-27.2\%}$	$87.27_{-0.0\%}$	$59.61_{-27.2\%}$	0.4647	48.93	0.8617
+ UNBIASED-MARK	$17.89_{+2.2\%}$	0.5508	$70.38_{-25.0\%}$	$88.86_{+1.8\%}$	$62.54_{-23.7\%}$	0.4855	19.93	0.5000
+ PROVABLE-MARK	$12.21_{-30.27\%}$	0.8020	$74.42_{-20.7\%}$	$96.70_{+10.8\%}$	$71.96_{-12.2\%}$	0.8796	10.21	0.9582
+ WATME$_{dictionary}$	$14.78_{-15.6\%}$	0.8044	$78.95_{-15.9\%}$	$97.43_{+11.6\%}$	$76.92_{-6.1\%}$	0.7897	10.96	0.9582
+ WATME$_{prompting}$	$16.22_{-7.4\%}$	0.7843	$69.65_{-25.8\%}$	$97.45_{-11.5\%}$	$67.87_{-17.2\%}$	0.7396	11.54	0.9519

Table 1: Performance comparison of Llama2-7B and Vicuna-v1.5-7B under different watermarking algorithms.