Can Watermarks Survive Translation? On the Cross-lingual Consistency of Text Watermark for Large Language Models

Zhiwei He, Binglin Zhou, Hongkun Hao, Aiwei Liu,
Xing Wang, Zhaopeng Tu, Zhuosheng Zhang, Rui Wang

Outline

Misuse of LLM

- Large language models (LLMs) have exhibited impressive content generation capabilities.
- Mitigating the misuse of LLM is important.
- Tagging and identifying LLM-generated content would help.

Text Watermark

- Text watermarking embeds a "message" into the LLM-generated content.

Text Watermark

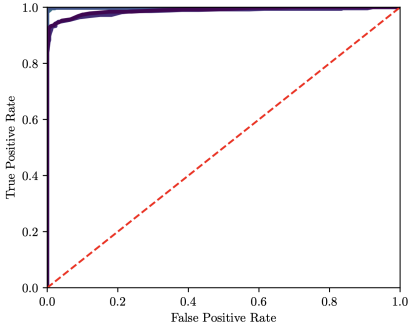
- Text watermarking embeds a "message" into the LLM-generated content.
  - invisible to human
  - can be detected algorithmically

Text Watermark

- Text watermarking embeds a "message" into the LLM-generated content.
  - invisible to human
  - can be detected algorithmically
- In the simplest form, the "message" can be a single bit indicating the presence of the watermark.

Text Watermark

- Text watermarking embeds a "message" into the LLM-generated content.
  - invisible to human
  - can be detected algorithmically
- In the simplest form, the "message" can be a single bit indicating the presence of the watermark.



ROC Curve | AUC=0.998
[KGW+23]

Can Watermarks Survive Translation?

- A malicious user could use a watermarked LLM to produce fake news in English, translate it into many other languages and spread it.
- The deceptive impact persists regardless of the language.

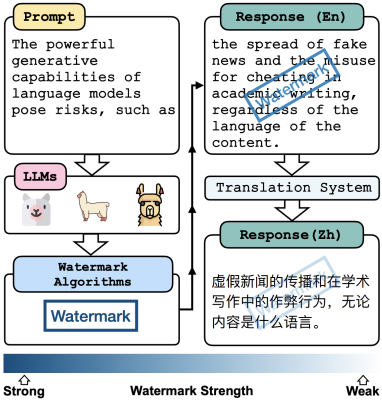


Figure 1: Illustration of watermark dilution in a cross-lingual environment. Best viewed in color.

Can Watermarks Survive Translation?

- A malicious user could use a watermarked LLM to produce fake news in English, translate it into many other languages and spread it.
- The deceptive impact persists regardless of the language.

Can watermarks survive translation?

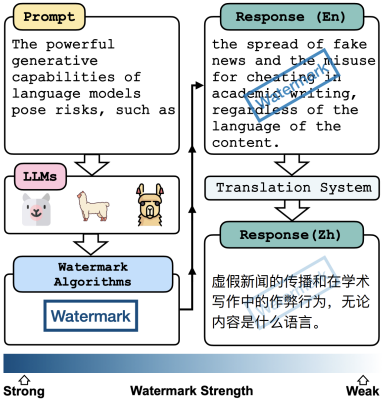
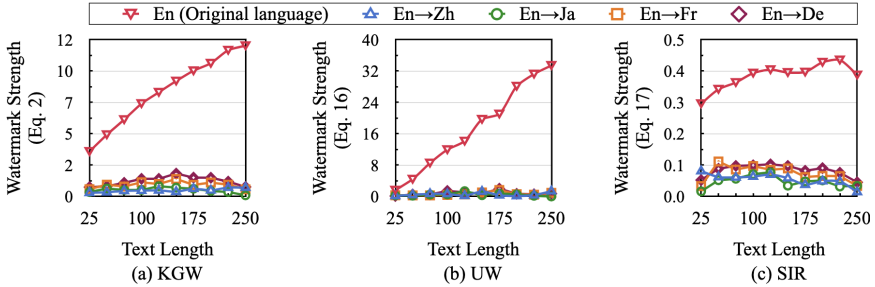


Figure 1: Illustration of watermark dilution in a cross-lingual environment. Best viewed in color.

Evaluation: Cross-lingual Consistency

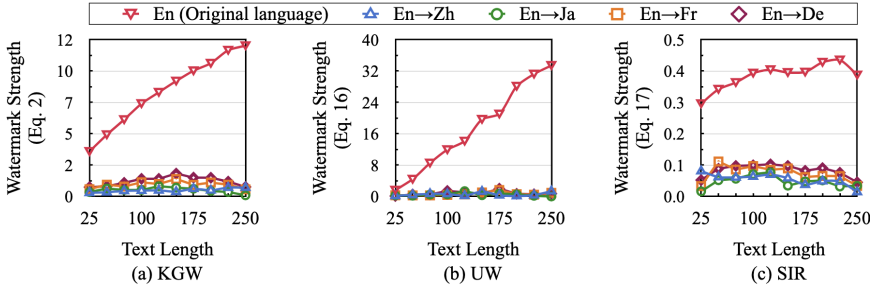
We define *cross-lingual consistency* to assess the ability of text watermarks to maintain their effectiveness after being translated into other languages.



[1] KGW:[KGW+23], UW: [HCW+23], SIR: [LPH+24]

Evaluation: Cross-lingual Consistency

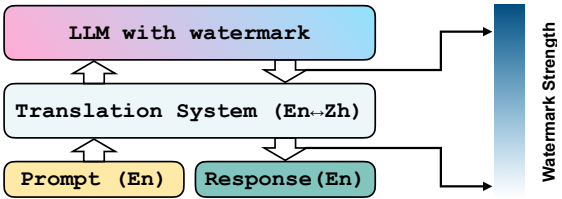
We define *cross-lingual consistency* to assess the ability of text watermarks to maintain their effectiveness after being translated into other languages.



Current text watermarking methods lack cross-lingual consistency.[1]

[1] KGW:[KGW+23], UW: [HCW+23], SIR: [LPH+24]

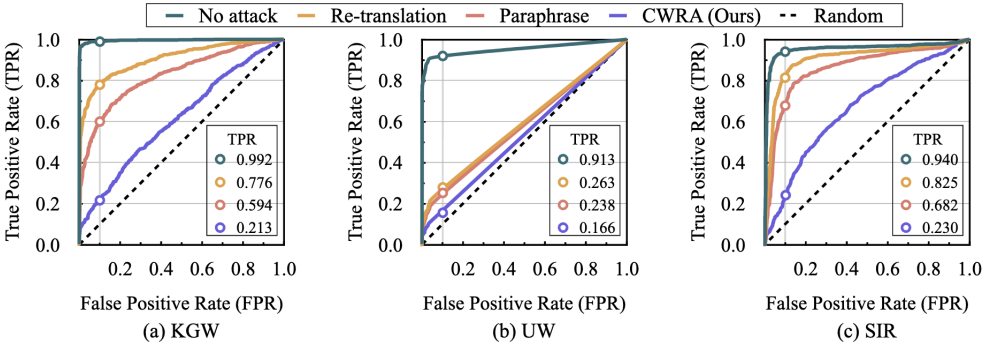
Cross-lingual Watermark Removal Attack (CWRA)



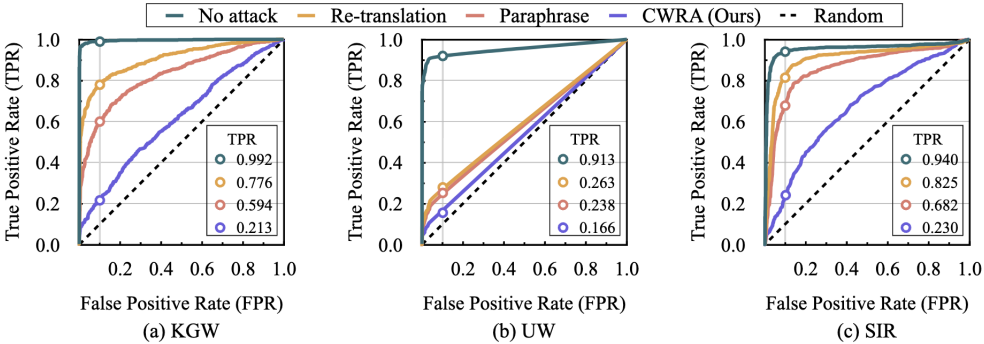
Cross-lingual Watermark Removal Attack (CWRA)

- CWRA wraps the query to the LLM into another language (Zh in the figure).
- The watermarks is diluted during the second translation step.

Performance: watermark detection



Performance: watermark detection



CWRA reduces AUC to a random-guessing level.

Performance: text quality

Attack \ WM	KGW			UW			SIR		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Text Summarization									
No attack	14.24	2.68	12.99	13.65	1.68	12.38	13.34	1.79	12.43
Re-translation	14.11	2.43	12.89	13.89	1.77	12.63	13.63	1.98	12.61
Paraphrase	15.10	2.49	13.69	14.72	1.95	13.31	15.56	2.11	14.14
CWRA (Ours)	**18.98**	**3.63**	**17.33**	**15.88**	**2.31**	**14.25**	**17.38**	**2.67**	**15.79**
Question Answering									
No attack	**19.00**	2.18	16.09	11.70	0.49	9.57	16.95	1.35	14.91
Re-translation	18.62	2.32	16.39	12.98	1.30	11.16	16.90	1.80	15.12
Paraphrase	18.45	2.24	**16.47**	14.38	1.37	13.07	17.17	1.79	**15.54**
CWRA (Ours)	18.23	**2.56**	16.27	**15.20**	**1.88**	**13.45**	**17.47**	**2.22**	15.53

Table 2: Comparative analysis of text quality impacted by different watermark removal attacks.

By choosing a pivot language in which the model excels, CWRA does not sacrifice text quality.

KGW-based watermarking

Vocab partition based on preceding text.

KGW-based watermarking

Vocab partition based on preceding text.

(1) compute a hash of $\boldsymbol{x}^{1:n}$: $h^{n+1} = H(\boldsymbol{x}^{1:n}) \cdots H(\cdot)$ can only use the last k tokens $\boldsymbol{x}^{n-k+1:n}$.

(2) seed a random number generator with h^{n+1} and randomly partitions $\mathcal{V}$ into two disjoint lists: the *green* list $\mathcal{V}_g$ and the *red* list $\mathcal{V}_r$,

(3) adjust the logits $\boldsymbol{z}^{n+1}$ by adding a constant bias δ ($\delta > 0$) for tokens in the green list:

$$\forall i \in \{1, 2, \ldots, |\mathcal{V}|\},$$

$$\tilde{\boldsymbol{z}}_i^{n+1} = \boldsymbol{z}_i^{n+1} + \Delta_i(\boldsymbol{x}^{1:n}) = \begin{cases} \boldsymbol{z}_i^{n+1} + \delta, & \text{if } v_i \in \mathcal{V}_g, \\ \boldsymbol{z}_i^{n+1}, & \text{if } v_i \in \mathcal{V}_r, \end{cases} \tag{1}$$

$$(\Delta \in \mathbb{R}^{|\mathcal{V}|}).$$

KGW-based watermarking

As a result, watermarked text will statistically contain more *green tokens*, an attribute unlikely to occur in human-written text.

Prompt
…The watermark detection algorithm can be made public, enabling third parties (e.g., social media platforms) to run it themselves, or it can be kept private and run behind an API. We seek a watermark with the following properties:
No watermark
Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words) Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet
With watermark
- minimal marginal probability for a detection attempt. - Good speech frequency and energy rate reduction. - messages indiscernible to humans. - easy for humans to verify.

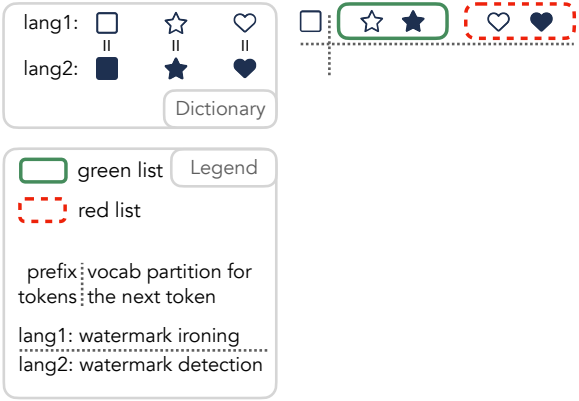
How to improve cross-lingual consistency?

- KGW-based watermarking methods fundamentally depend on the partition of the vocab, i.e., the red and green lists.

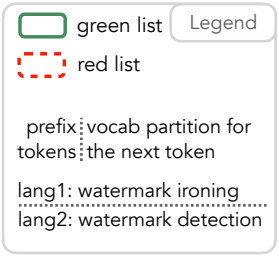
Cross-lingual consistency

the green tokens in the watermarked text will still be recognized as green tokens after being translated into other languages

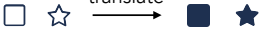
A simple case study



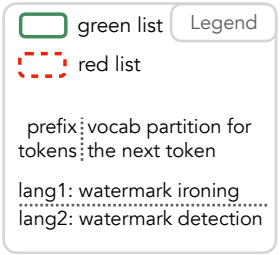
A simple case study

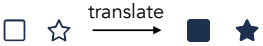


A simple case study



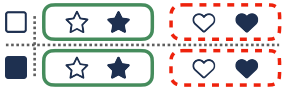
A simple case study



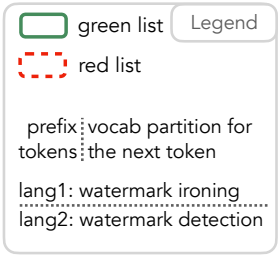


- What conditions must the vocabulary partition satisfy so that the ★, the semantic equivalent of the ☆, is also included in the green list?

A simple case study

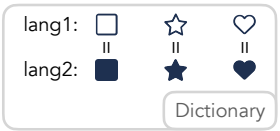


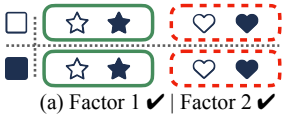
(a) Factor 1 ✔ | Factor 2 ✔

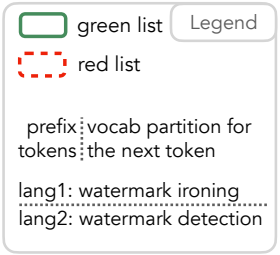


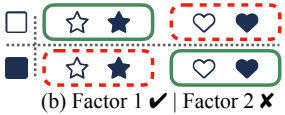
- **Factor 1**: semantically similar tokens should be in the same list (either red or green)
- **Factor 2**: the vocab partitions for semantically similar prefixes should be the same.

A simple case study



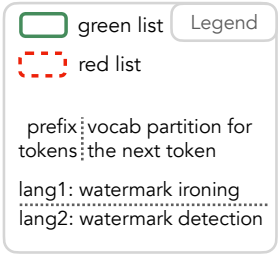


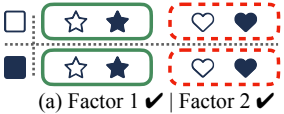


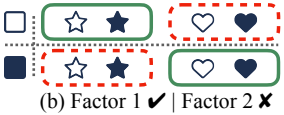


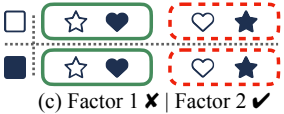
- **Factor 1**: semantically similar tokens should be in the same list (either red or green)
- **Factor 2**: the vocab partitions for semantically similar prefixes should be the same.

A simple case study



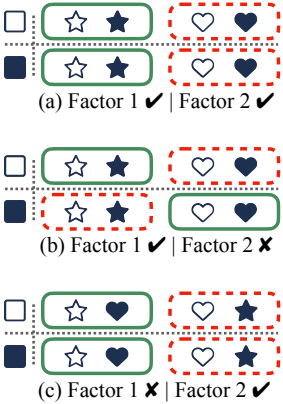

(a) Factor 1 ✔ | Factor 2 ✔


(b) Factor 1 ✔ | Factor 2 ✘


(c) Factor 1 ✘ | Factor 2 ✔

- **Factor 1**: semantically similar tokens should be in the same list (either red or green)
- **Factor 2**: the vocab partitions for semantically similar prefixes should be the same.

A simple case study



- **Factor 1**: semantically similar tokens should be in the same list (either red or green)
- **Factor 2**: the vocab partitions for semantically similar prefixes should be the same.

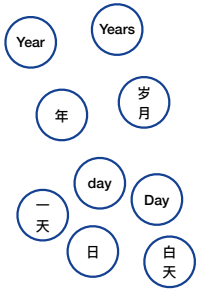
Factor 1 & 2 must be satisfied simultaneously.

Defense Method: X-SIR

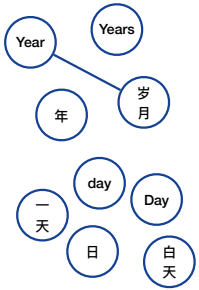
- **Factor 1**: semantically similar tokens should be in the same list (either red or green)
- **Factor 2**: the vocab partitions for semantically similar prefixes should be the same.

Fortunately, SIR [LPH+24] has already optimized for the **Factor 2**.
Based on SIR, we discuss how to achieve the **Factor 1** and name our method X-SIR.

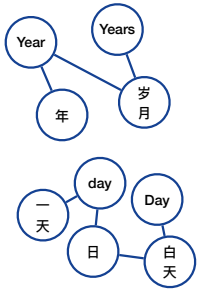
Defense Method: X-SIR



1. Initialize a graph where each node is a token of the model vocabulary.

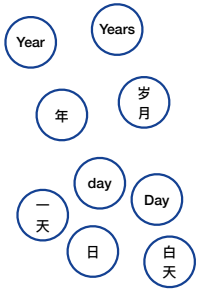


2. If the two tokens are an entry in an external dictionary, add an edge between the two nodes.

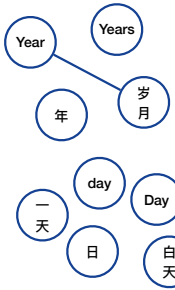


3. Each connected component is a cluster of semantically similar tokens

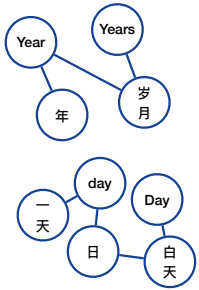
Defense Method: X-SIR



1. Initialize a graph where each node is a token of the model vocabulary.



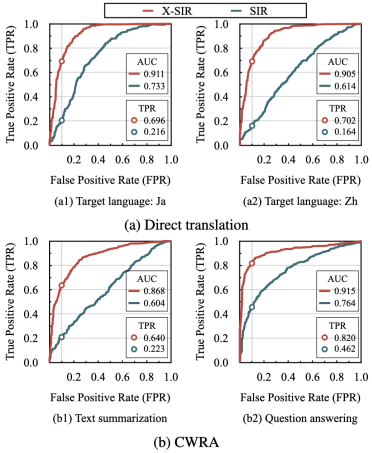
2. If the two tokens are an entry in an external dictionary, add an edge between the two nodes.



3. Each connected component is a cluster of semantically similar tokens

Vocabulary partition: token-level $\Rightarrow$ cluster-level

Performance: watermark detection



- AUC: $+0.20$
- TPR: $+0.40$

Performance: text quality

Method	ROUGE-1	ROUGE-2	ROUGE-L
	Text Summarization		
SIR	13.34	1.79	12.43
X-SIR	**15.65**	**2.04**	**14.29**
	Question Answering		
SIR	**16.95**	1.35	**14.91**
X-SIR	16.77	**1.39**	14.07

Table 4: Effects of X-SIR and SIR on text quality.

Summary

A closed-loop study:

- **Evaluation**: We reveal the deficiency of current text watermarking technologies in maintaining cross-lingual consistency.
- **Attack**: Based on this finding, we propose CWRA that successfully bypasses watermarks without degrading the text quality.
- **Defense**: We identify two key factors for improving cross-lingual consistency and propose X-SIR as a defense method against CWRA.

Paper & Code

https://arxiv.org/abs/2402.14007

https://github.com/zwhe99/X-SIR

Bibliography I

[HCW+23] Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang, Unbiased watermark for large language models, arXiv preprint arXiv:2310.10669 (2023).

[KGW+23] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein, A watermark for large language models, Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 202, PMLR, 23–29 Jul 2023, pp. 17061–17084.

[LPH+24] Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen, A semantic invariant robust watermark for large language models, The Twelfth International Conference on Learning Representations, 2024.