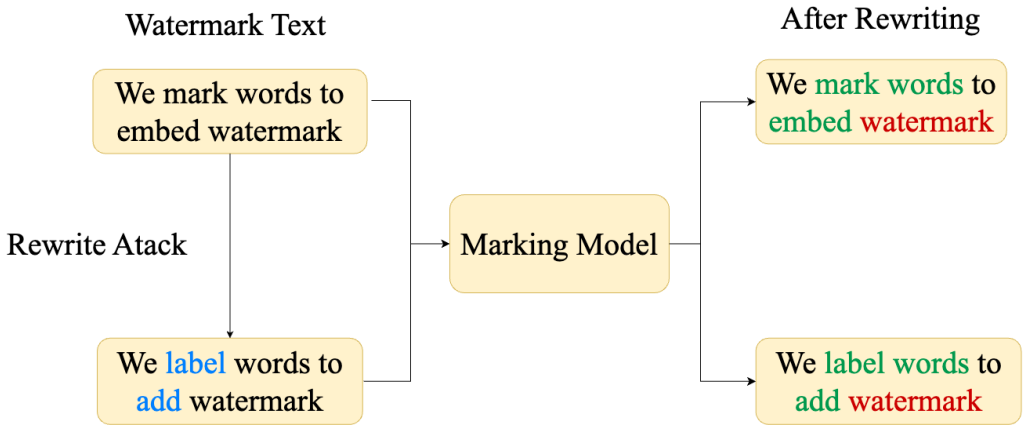
Semantic-oriented Robust Text Watermarking for Large Language Models

Cheng SONG, Kun ZHANG, Bo LI, Hailin ZHOU, Richang HONG

Frontiers of Computer Science, DOI: 10.1007/s11704-026-50763-3

Problems & Ideas

- Problems of conventional text watermarks:
  - Existing methods mark words based on word-level hashing that exhibit an overdependence on individual word.
  - Rewriting attack can compromise the robustness of the watermark by altering the hash values of certain words in the text.
- Ideas: Marking word with semantics which are resistant to rewriting attacks, enhanced watermark robustness will be achieved.



The proportion of marked words in a watermarked text can be extracted by processing it through the same marking model used during embedding. As the mark is semantically based, this proportion remains constant even after rewriting attack.

Main Contributions

- Contributions:
  - Evaluation using the F1-score and AUC shows that the proposed method significantly enhances robustness against both rewriting and re-translation attacks.
  - This approach effectively embedding of a high-robustness watermark with the preservation of textual semantic quality.

Dataset	Algorithm	No attack			Rewrite			Re-translate		
		F1↑ 5%FPR	F1↑ 1%FPR	AUC↑	F1↑ 5%FPR	F1↑ 1%FPR	AUC↑	F1↑ 5%FPR	F1↑ 1%FPR	AUC↑
HC3	ContextLS	97.0%	97.1%	99.8%	0.5%	0.2%	14.0%	0.5%	-	22.9%
	RMNLW	93.5%	79.6%	98.3%	1.0%	0.2%	26.6%	3.3%	0.5%	34.2%
	WTBLM	96.9%	98.2%	99.6%	9.1%	2.4%	52.9%	7.1%	1.5%	49.0%
	RTW (Ours)	**97.6%**	**99.1%**	**99.8%**	**17.2%**	**5.8%**	**66.0%**	**68.5%**	**54.9%**	**87.8%**
bloomz	ContextLS	97.1%	98.2%	99.8%	0.2%	-	11.7%	-	-	15.5%
	RMNLW	93.3%	79.9%	98.4%	1.7%	0.2%	25.6%	0.5%	-	26.4%
	WTBLM	97.6%	98.9%	**100.0%**	31.3%	9.7%	66.2%	62.4%	35.4%	82.2%
	RTW (Ours)	**98.5%**	**99.4%**	**100.0%**	**61.4%**	**42.5%**	**86.0%**	**78.1%**	**61.5%**	**91.5%**
chatGPT	ContextLS	97.0%	97.9%	99.7%	-	-	13.8%	0.2%	-	15.9%
	RMNLW	94.0%	80.1%	98.3%	1.0%	-	24.4%	1.0%	0.2%	26.1%
	WTBLM	98.3%	99.3%	**100.0%**	28.2%	12.1%	68.1%	57.7%	39.5%	83.3%
	RTW (Ours)	**98.9%**	**99.9%**	**100.0%**	**62.5%**	**42.0%**	**86.5%**	**74.3%**	**59.5%**	**90.6%**
cohere	ContextLS	97.5%	98.4%	99.9%	-	-	11.8%	-	-	14.5%
	RMNLW	93.3%	67.7%	98.2%	1.0%	-	23.7%	0.9%	-	26.3%
	WTBLM	98.2%	99.3%	**100.0%**	28.9%	12.7%	67.3%	58.2%	38.1%	82.1%
	RTW (Ours)	**100.0%**	**100.0%**	**100.0%**	**64.4%**	**43.3%**	**86.9%**	**76.9%**	**60.0%**	**91.1%**
davinci	ContextLS	97.2%	98.1%	99.8%	-	-	9.8%	0.5%	0.2%	13.8%
	RMNLW	86.7%	67.1%	97.7%	0.2%	-	22.0%	1.7%	0.2%	25.7%
	WTBLM	97.6%	98.6%	99.9%	32.5%	9.2%	68.2%	56.8%	28.6%	83.0%
	RTW (Ours)	**97.7%**	**99.8%**	**100.0%**	**61.6%**	**43.0%**	**86.7%**	**74.0%**	**60.0%**	**90.9%**
dolly	ContextLS	97.0%	96.9%	99.6%	-	-	11.9%	0.2%	-	14.8%
	RMNLW	96.5%	82.6%	99.2%	1.4%	0.5%	29.0%	2.1%	0.2%	28.7%
	WTBLM	97.1%	98.0%	99.7%	25.1%	7.6%	70.0%	44.7%	23.2%	78.1%
	RTW (Ours)	**97.7%**	**99.9%**	**100.0%**	**61.7%**	**31.1%**	**87.3%**	**72.9%**	**47.7%**	**89.0%**
flant5	ContextLS	97.3%	98.4%	99.7%	-	-	12.2%	-	-	14.3%
	RMNLW	87.3%	65.0%	97.9%	1.0%	0.2%	24.6%	1.0%	0.2%	27.3%
	WTBLM	97.7%	98.9%	99.9%	22.4%	9.4%	66.2%	51.4%	30.1%	81.3%
	RTW (Ours)	**97.8%**	**99.6%**	**100.0%**	**60.5%**	**42.2%**	**85.4%**	**74.0%**	**60.4%**	**90.7%**

	HC3	bloomz	chatGPT	cohere	davinci	dolly	flant5
ContextLS	0.998	0.996	0.997	0.997	0.997	0.998	0.997
RMNLW	0.998	0.997	0.998	0.998	0.998	0.998	0.998
WTBLM	0.996	0.989	0.999	0.999	0.999	0.999	0.999
RTW(Ours)	0.981	0.926	0.927	0.929	0.927	0.948	0.926

The results evaluated on multiple public datasets. Left:model performance against various attacks ; Right: average similarity between watermark and original text.