Provable Robust Watermarking for AI-Generated Text

Xuandong Zhao, Prabhanjan Ananth, Lei Li, Yu-Xiang Wang
{xuandongzhao, prabhanjan, leili, yuxiangw}@cs.ucsb.edu

SCAN ME

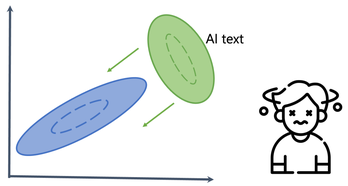
Motivation

- **Potential harms of LLM**
  - Generate fake news
  - Contaminate web content
  - Assist in academic dishonesty
- **If most text in daily life is AI generated?**

Detect AI-generated text

Distinguish AI-generated text from human

Turing-test



AI-generated text increasingly resembles human-generated text

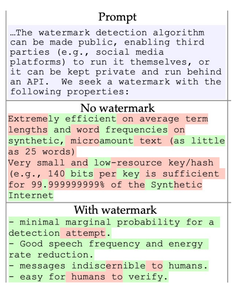
- **ZeroGPT/GPTZero/DetectGPT**
  - Not robust to distribution changes
  - Prone to biases
  - Vulnerable to adversarial attacks

Watermarking digital text

PROVE VS PREDICT

Watermarking

AI Classifier



GPTWatermark

- **Watermarking text generation**
  1. Randomly generate a watermark key k. Use watermark key to partition the vocabulary into a **Green List** of size $\gamma|V|$
  2. For $t = 1, 2, \ldots$
     1. Apply the language model to prior tokens to obtain a logit vector ℓ_t
     2. **Add δ to each green list logit**. Apply the Softmax operator

$$\hat{\mathbf{p}}_t[v] = \begin{cases} \frac{\exp(\boldsymbol{\ell}_t[v]+\delta)}{\sum_{i\in Red}\exp(\boldsymbol{\ell}_t[i])+\sum_{i\in Green}\exp(\boldsymbol{\ell}_t[i]+\delta)}, & v \in Green \\ \frac{\exp(\boldsymbol{\ell}_t[v])}{\sum_{i\in Red}\exp(\boldsymbol{\ell}_t[i])+\sum_{i\in Green}\exp(\boldsymbol{\ell}_t[i]+\delta)}, & v \in Red. \end{cases}$$

     3. Decode the next token using the watermarked distribution $\hat{\mathbf{p}}_t$
- **Watermarking text detection**
  1. Use the watermark detection key k to find the Green List
  2. Tokenize the suspect text and calculate the number of green list tokens $|\boldsymbol{y}|_G = \sum_{t=1}^{n} \mathbf{1}(y_t \in G)$
  3. Assume the null hypothesis is H_0: *The text sequence is generated with no knowledge of the green list rule.* Compute the z-statistic for this test:

$$z = \left(|\boldsymbol{y}|_G - \gamma n\right) / \sqrt{n\gamma(1-\gamma)}$$

  4. If $z > \tau$, the suspect text is watermarked

Theoretical framework

- **ω-Quality of watermarked output**

$$D\left(\hat{\mathbf{p}}_t \| \mathbf{p}_t\right) \leq \omega$$

- **α-Type I error ("No false positives"):**

$$\mathbb{P}\left[\text{Detect}(\text{k}, \boldsymbol{y}) = 1 \; ; \; \substack{(\hat{\mathcal{M}},\text{k})\sim\text{Watermark}(\mathcal{M}) \\ \boldsymbol{y}\sim\mathcal{A}(\boldsymbol{x},\text{aux})}\right] \leq \alpha_{(\boldsymbol{x},\mathcal{M})}.$$

- **β-Type II error ("No false negatives"):**

$$\mathbb{P}\left[\text{Detect}(\text{k}, \boldsymbol{y}) = 0 \; ; \; \substack{(\hat{\mathcal{M}},\text{k})\sim\text{Watermark}(\mathcal{M}) \\ \boldsymbol{y}\sim\hat{\mathcal{M}}(\boldsymbol{x})}\right] \leq \beta_{(\boldsymbol{x},\mathcal{M})}.$$

- **Security property**

The adversary needs to make enough edits to evade detection.

Robustness property

- **GPTWatermark robustness to editing**

Twice the robustness!

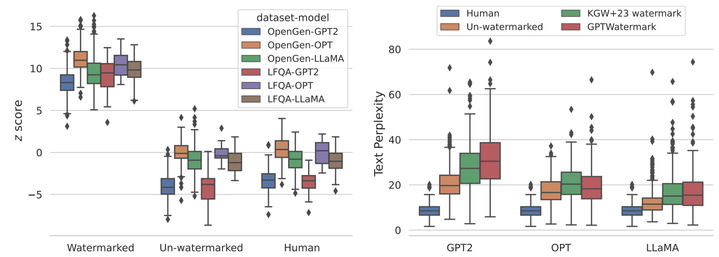
$$z_{\boldsymbol{u}} \geq z_{\boldsymbol{y}} - \max\{\frac{(1+\gamma/2)\eta}{\sqrt{n}}, \frac{(1-\gamma/2)\eta}{\sqrt{n-\eta}}\}.$$

- **KGW⁺23 watermark robustness to editing**

$$z_{\boldsymbol{u}} \geq z_{\boldsymbol{y}} - \max\{\frac{(2+\gamma/2)\eta}{\sqrt{n}}, \frac{(2-\gamma/2)\eta}{\sqrt{n-\eta}}\}.$$

Experiment results

- **z-score and text perplexity**



- **Robustness against paraphrasing attack**

Setting	Method	OpenGen 1% FPR TPR	OpenGen 1% FPR F1	OpenGen 10% FPR TPR	OpenGen 10% FPR F1	LFQA 1% FPR TPR	LFQA 1% FPR F1	LFQA 10% FPR TPR	LFQA 10% FPR F1
No attack	KGW+23	1.000	0.995	1.000	0.952	1.000	0.995	1.000	0.952
	GPTWatermark	1.000	0.995	1.000	0.952	1.000	0.995	1.000	0.952
ChatGPT	KGW+23	0.565	0.704	0.853	0.747	0.327	0.453	0.673	0.490
	GPTWatermark	0.866	0.910	0.961	0.818	0.442	0.568	0.865	0.584
DIPPER-1	KGW+23	0.386	0.546	0.738	0.720	0.372	0.534	0.740	0.767
	GPTWatermark	0.729	0.830	0.922	0.837	0.639	0.770	0.909	0.865
DIPPER-2	KGW+23	0.490	0.646	0.810	0.769	0.432	0.595	0.845	0.839
	GPTWatermark	0.777	0.862	0.941	0.852	0.693	0.810	0.948	0.894
BART	KGW+23	0.342	0.505	0.667	0.759	0.457	0.617	0.783	0.836
	GPTWatermark	0.590	0.730	0.861	0.857	0.656	0.784	0.885	0.897

- **Distinguishing human-written text**