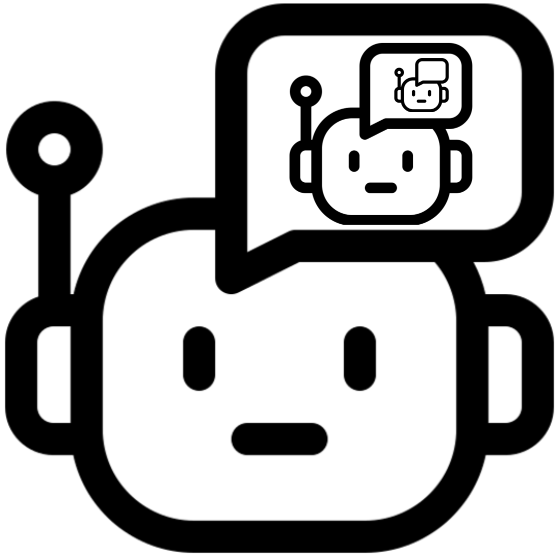
Provable Robust Watermarking for AI-Generated Text

Xuandong Zhao, Prabhanjan Ananth, Lei Li, Yu-Xiang Wang

xuandongzhao@ucsb.edu

COMPUTER SCIENCE
UC SANTA BARBARA
Computing. ReInvented.

Why do we need to detect AI-generated text?



Model Degeneration

Can you distinguish human vs. machine generated text?

Through the town, and past the lights,
Oh, how the bells do ring!
They chime with glee
For you and me
As carols we joyfully sing.

Machine

Over the river, and through the wood,
Oh, how the wind does blow!
It stings the toes
And bites the nose
As over the ground we go.

Human

Child, Lydia Maria. "Thanksgiving Day." 1844.

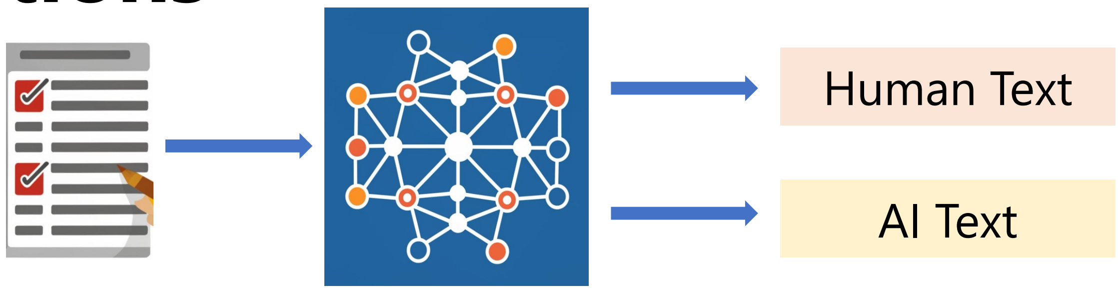
How to detect AI-generated text?

1. Prefix: "As a large language model…"

→ trivial to remove from text!

2. Database of all completions

→ privacy?



3. Train classification models [GPTZero, Turnitin, …]

→ too many false positives, not robust to OOD?

Watermarking is a promising solution!

Plant subtle but distinctive signals deliberately within the content to enable downstream detection

Watermarking vs. AI Classifier

Active | **Passive**

Watermark have a long history

The *Crown CA* watermark found on many British Commonwealth stamps

https://en.wikipedia.org/wiki/Watermark

Desired Properties of an Ideal Watermark

- **Quality of Generated Text**
- **Detection Accuracy Guarantee**
  - Type I error: "No false positives" → won't catch human text
  - Type II error: "No false negatives" → won't miss LLM text
- **Robustness Guarantee**
  - Be robust against evasion attacks, e.g., post-editing.

We develop Unigram-Watermark and the first theoretical framework for LLM watermarking

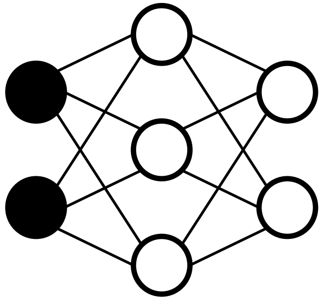
- **Quality Guarantee**
  - Watermarked LLM and original LLM are **indistinguishable.**
- **Detection Guarantees**
  - Type I error → 0 **exponentially** as text gets larger.
  - Type II error → 0 **exponentially** as text gets larger (under natural technical conditions).
- **Provably Robust** to Edits -- **Twice as robust** as the previous method.

Kirchenbauer et al. A Watermark for Large Language Models, ICML 2023

Revisit the Language Model

P(next word y_t | Prompt x, previous words $y_{1:t-1}$)

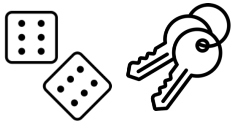
"Santa Barbara has nice ___"



	Logits	Probability
weather	2.24	0.414
beaches	1.73	0.329
eyes	1.01	0.249
shoes	-1.98	0.006
⋮		

The **universe of words** is called a **vocabulary** V

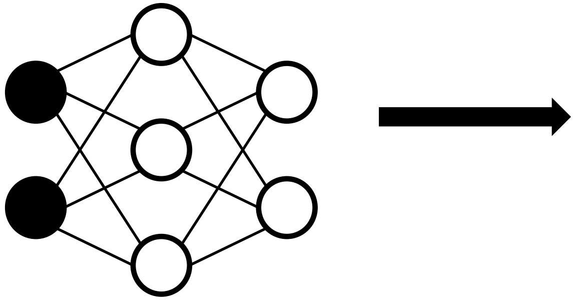
Unigram-Watermark



Random split

Green List	Red List
Santa	Barbara
beach	weather
eye	shoes
…	…

"Santa Barbara has nice ___"



	Logits	Perturb	Probability
weather	2.24		0.131
beaches	1.73	$+\delta$	0.581
eyes	1.01	$+\delta$	0.249
shoes	-1.98		0.006
⋮			

Increase the probability of green tokens slightly.

Decrease the probability of red tokens slightly.

X. Zhao, P. Ananth, L. Li, YX Wang. Provable Robust Watermarking for AI-Generated Text, ICLR 2024

Unigram-Watermark: Adding Watermark

1. Randomly split the vocabulary into a Green List ($\gamma|V|$) and a Red List ($(1-\gamma)|V|$)
2. For $t = 1, 2, \ldots$
   1. Get logit vector ℓ_t from LLM
   2. Add δ to each green list logit and apply Softmax

$$\hat{p}_t = \mathrm{softmax}(\ell_t + \delta \cdot \mathbf{1}(v_t \in Green))$$

   3. Generate next token using $\hat{p}_t$

Unigram-Watermark: Detecting Watermark

Input: Suspect text $y = [y_1, \ldots, y_n]$, e.g. "Over the …"

1. Compute the ***z*-score**:

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

Num of Green tokens

2. If $z >$ *threshold* then

    Return "y is watermarked"

  Else

    Return "no evidence"

Unigram-Watermark Examples

LLaMA-13B, unwatermarked
z-score=-2.4

Prompt: Can I succeed after many failures?

A: Of course it is, and that is how we improve. Saying "I can\'t do that" is never a good thing. Sometimes we think we\'ve tried all we can and that "isn\'t enough". That is the time when we ask for help. The root of all evils is to be a secret. Honesty and self-criticism is necessary for improvement. The measure of intelligence is the ability to change. [continues...]

LLaMA-13B, watermarked
z-score=11

Prompt: Can I succeed after many failures?

A: When most people are confronted with failure, they cannot imagine such a thing happening. When one faces business reverses and bankruptcy, it seems impossible. When we are rejected it looks as if we are going to be rejected forever. However, it does not need to be this way. The human spirit simply will not give up. [continues...]

Theoretical Contributions

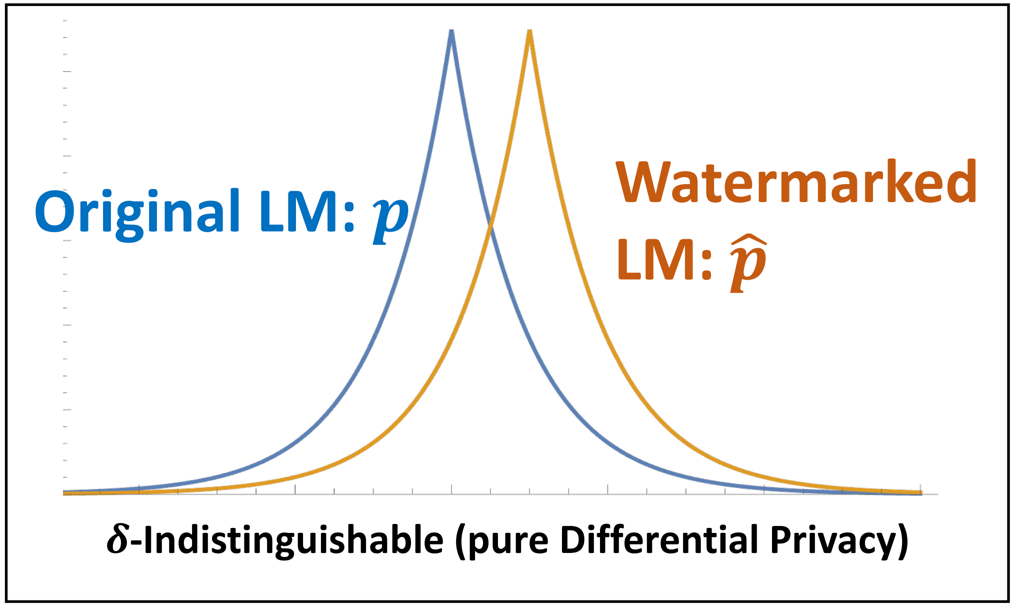
Quality Guarantee

Theorem: Any prompt, any prefix text. Any Renyi-Divergence $D_\alpha(p||\hat{p}) \leq \min\{\delta, \frac{\alpha\delta^2}{8}\}$

$\alpha = 1$

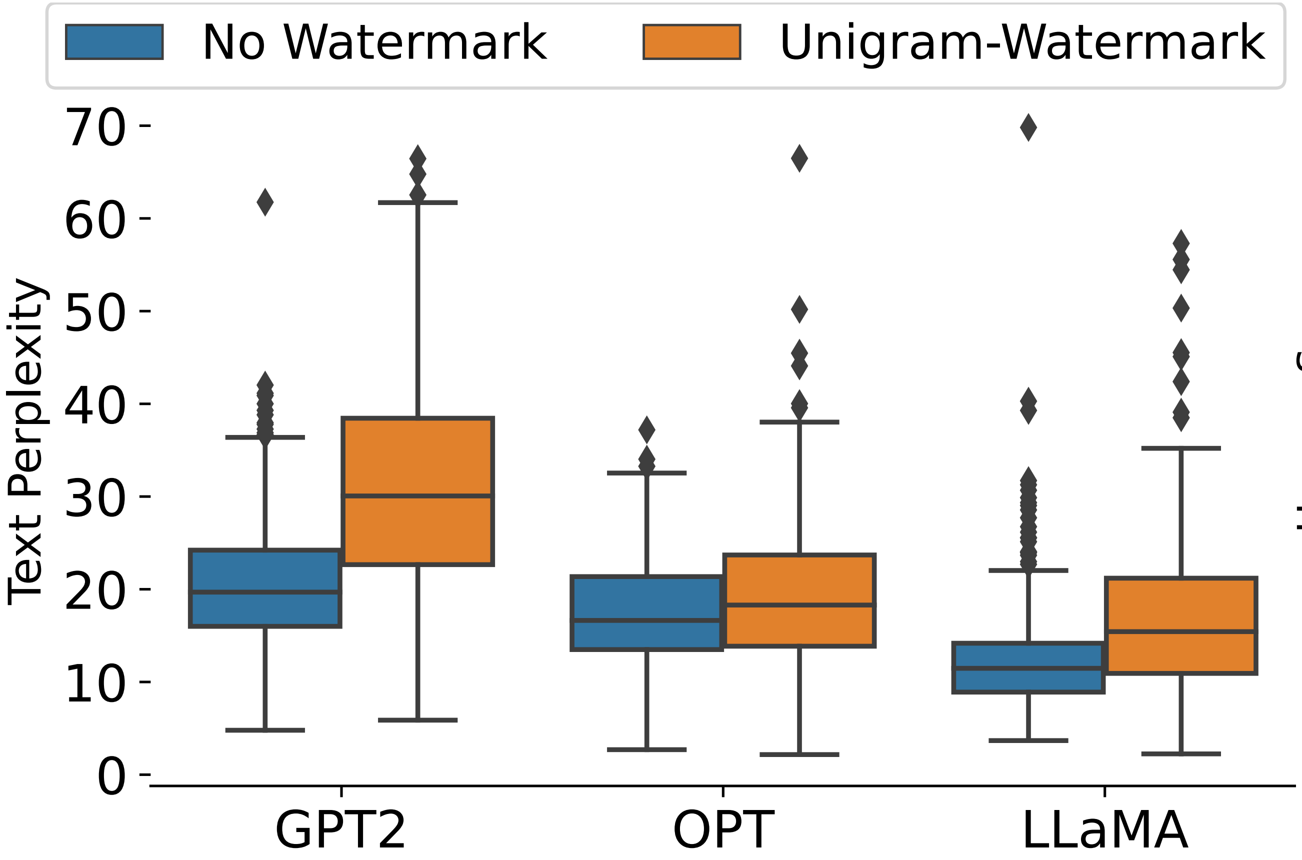
⬇

KL-divergence

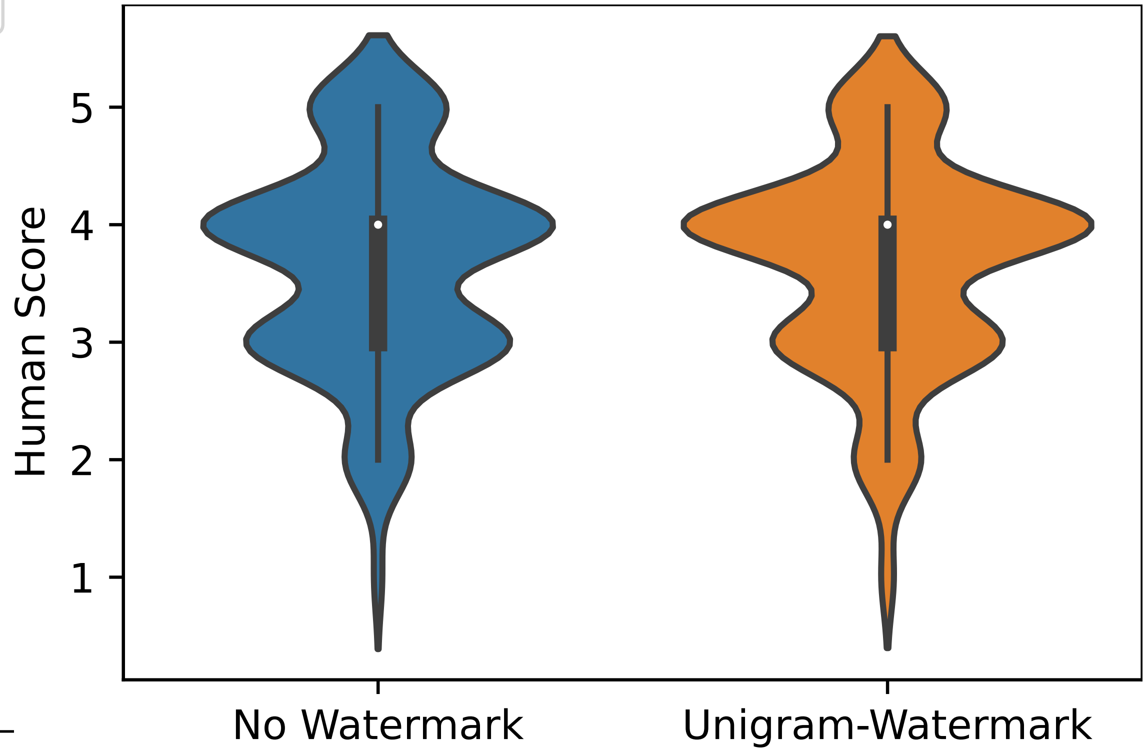


X. Zhao, P. Ananth, L. Li, YX Wang. Provable Robust Watermarking for AI-Generated Text, ICLR 2024

The performance of the watermarked LLM remains strong!



GPT3 PPL Score



Human Evaluation

X. Zhao, P. Ananth, L. Li, YX Wang. Provable Robust Watermarking for AI-Generated Text, ICLR 2024

Detection Guarantee: Type I/II Error

Theorem (informal):
If the suspect text $\boldsymbol{y}$ is independent to the secret key (i.e., the green list):

$$z_{\boldsymbol{y}} \asymp O(1)$$

No False Positive

If the suspect text $\boldsymbol{y}$ is generated using watermarked LM:

$$z_{\boldsymbol{y}} \asymp \delta\sqrt{n}$$

Only True Positive

Unigram-Watermark is robust to edits!

Theorem: Adversary take watermarked output y, Adversary edits to get to a new text u. If Edit Distance $ED(y, u) \leq \eta$, then

Unigram-Watermark is robust up to $O(n)$ arbitrary edits

Comparing to the Previous Watermark

Unigram-Watermark is provably **2x as robust** to edits (deletions, replacements, additions) compared to **KGW-Watermark**.

Kirchenbauer et al. A Watermark for Large Language Models, ICML 2023 (Outstanding Paper Award)

Experiment

- Two long-form text datasets
  - **OpenGen**: 3K chunks sampled from WikiText-103
  - **LFQA**: long-form question-answering dataset from Reddit

- Three state-of-the-art public language models
  - **GPT2-XL-1.5B** [Radford et al., 2019]
  - **OPT-1.3B** [Zhang et al., 2022]
  - **LLaMA-7B** [Touvron et al., 2023]

Paraphrasing Attacks

Generated text with Unigram-Watermark (LLaMA-7B)

Prompt: "Rewrite the following paragraph:"

ChatGPT

Paraphrasing model:

DIPPER
[Krishana et al, 2023]

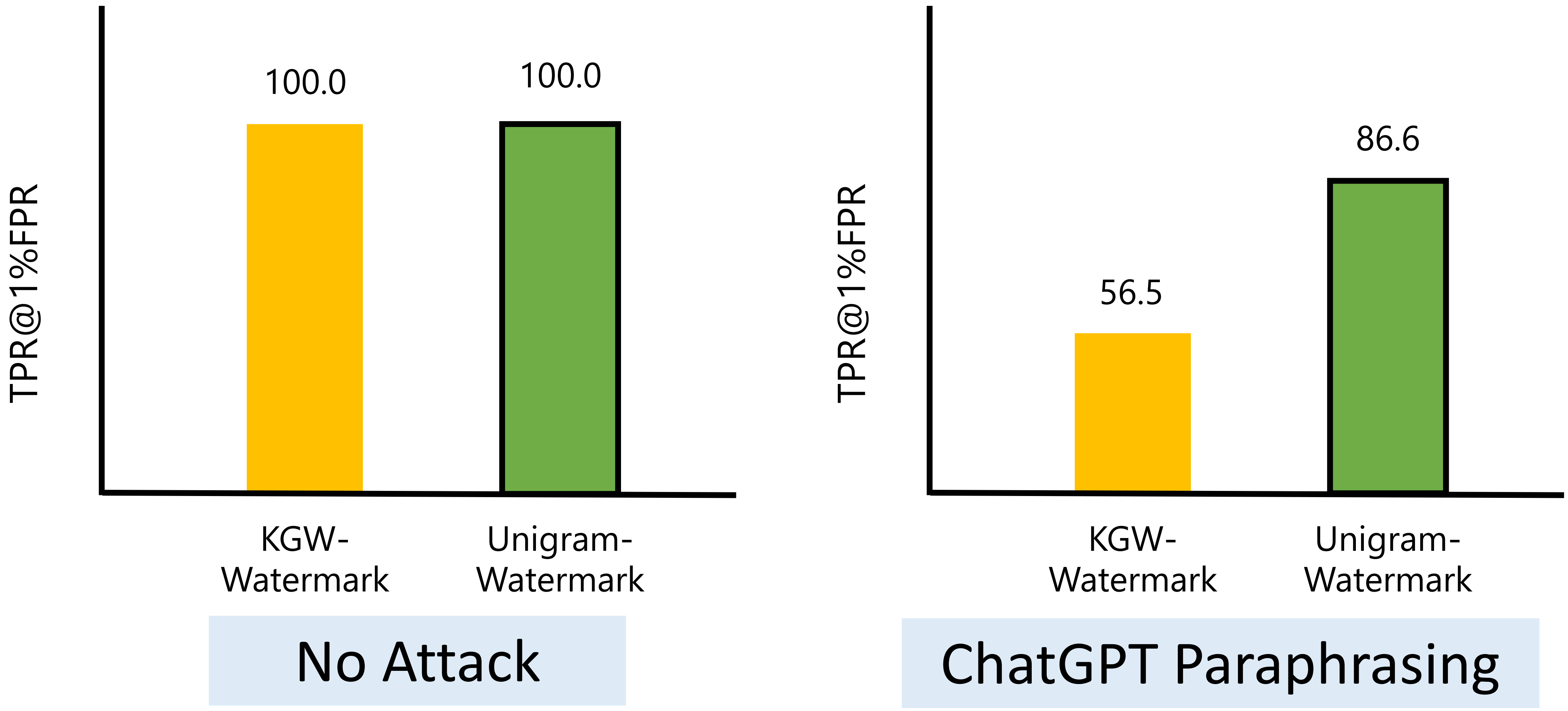
Summarization model:

BART
[Lewis et al, 2019]

Adversary wants to evade the detection

X. Zhao, P. Ananth, L. Li, YX Wang. Provable Robust Watermarking for AI-Generated Text, ICLR 2024

Robustness Against Paraphrasing Attack

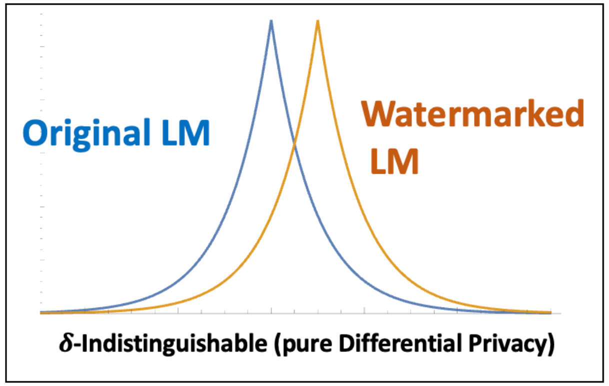


X. Zhao, P. Ananth, L. Li, YX Wang. Provable Robust Watermarking for AI-Generated Text, ICLR 2024

Unigram-Watermark is Accurate and Provably Robust

Quality

Watermarked LLM and original LLM are **indistinguishable**.



Detection

As n gets larger

↓

False Positive Rate
False Negative Rate

↓

Exponentially decrease to 0

Robust

Provably robust to edits: Twice as robust as notable baseline. [Kirchenbauer et al. 2023]

X. Zhao, P. Ananth, L. Li, YX Wang. Provable Robust Watermarking for AI-Generated Text, ICLR 2024