Watermark for language models

Ponomarchuk Anna

A **watermark** is a hidden pattern in text that is imperceptible to humans, while making the text algorithmically identifiable as synthetic.

A **watermark** is a hidden pattern in text that is imperceptible to humans, while making the text algorithmically identifiable as synthetic.

Watermark properties:

- can be algorithmically detected without any knowledge of the model parameters

A **watermark** is a hidden pattern in text that is imperceptible to humans, while making the text algorithmically identifiable as synthetic.

Watermark properties:

- can be algorithmically detected without any knowledge of the model parameters
- text can be generated using a standard language model without re-training

A **watermark** is a hidden pattern in text that is imperceptible to humans, while making the text algorithmically identifiable as synthetic.

Watermark properties:

- can be algorithmically detected without any knowledge of the model parameters
- text can be generated using a standard language model without re-training
- is detectable from only a contiguous portion of the generated text

A **watermark** is a hidden pattern in text that is imperceptible to humans, while making the text algorithmically identifiable as synthetic.

Watermark properties:

- can be algorithmically detected without any knowledge of the model parameters
- text can be generated using a standard language model without re-training
- is detectable from only a contiguous portion of the generated text
- cannot be removed without modifying a significant fraction of the generated tokens

A **watermark** is a hidden pattern in text that is imperceptible to humans, while making the text algorithmically identifiable as synthetic.

Watermark properties:

- can be algorithmically detected without any knowledge of the model parameters
- text can be generated using a standard language model without re-training
- is detectable from only a contiguous portion of the generated text
- cannot be removed without modifying a significant fraction of the generated tokens
- it is possible to compute a rigorous statistical measure of confidence that the watermark has been detected

Just a friendly reminder

- vocabulary V contains tokens
- $s^{(-N_p)}$,…, $s^{(-1)}$ are tokens from the prompt of length N_p
- $s^{(0)}$,…, $s^{(T)}$ are tokens generated by an AI system in response to the prompt

A language model for next word prediction, is a function f that accepts as input a sequence of known tokens $s^{(-N_p)}$,…, $s^{(t-1)}$, and then outputs a vector of $|V|$ logits that are passed through a softmax operator to convert them into a discrete probability distribution over the vocabulary.

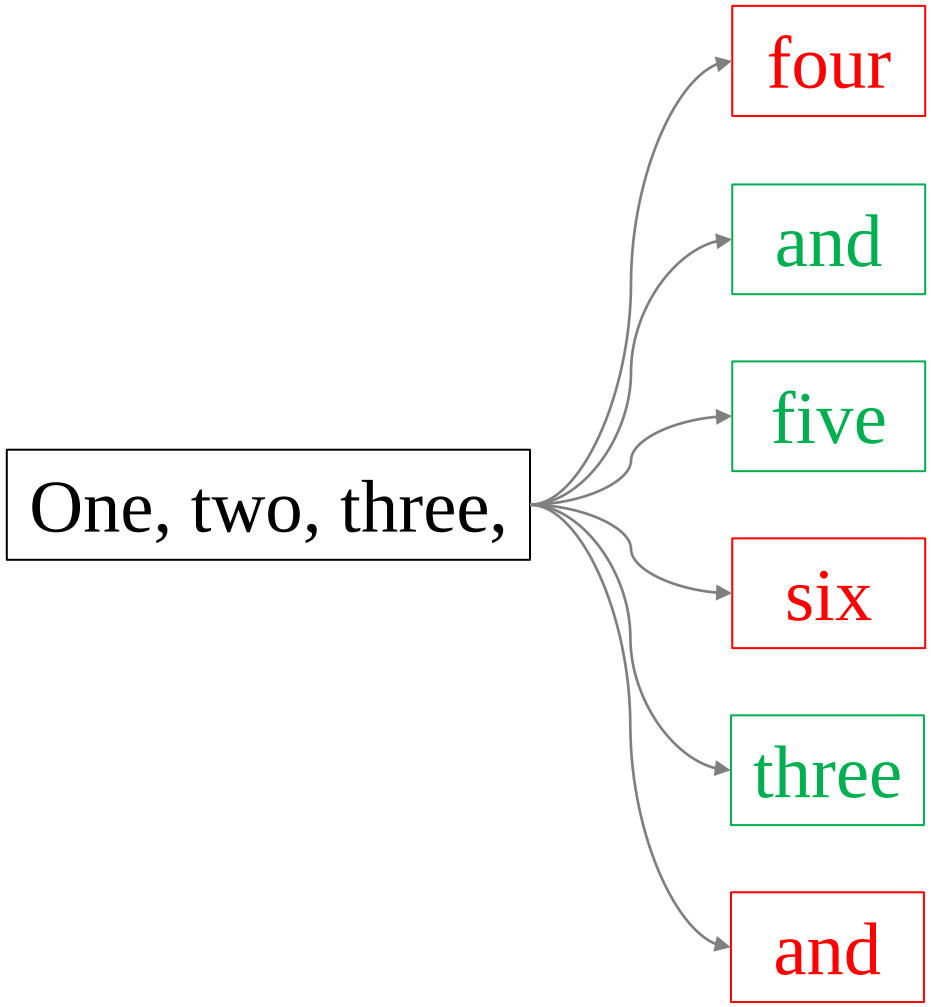
The difficulty of watermarking low-entropy sequences

Без труда не выловишь и
рыбку из пруда

for(i=0;i<n;i++) sum+=array[i]

The first few tokens strongly determine the following tokens.

«Hard» watermark algorithm



«Hard» watermark algorithm

Input: prompt, $s^{(-N_p)} \ldots s^{(-1)}$
for $t = 0, 1, \cdots$ **do**

1. Apply the language model to prior tokens $s^{(-N_p)} \ldots s^{(t-1)}$ to get a probability vector $p^{(t)}$ over the vocabulary.

2. Compute a hash of token $s^{(t-1)}$, and use it to seed a random number generator.

3. Using this seed, randomly partition the vocabulary into a "green list" G and a "red list" R of equal size.

4. Sample $s^{(t)}$ from G , never generating any token in the red list.

end for

Detecting the watermark

H_0: The text sequence is generated with no knowledge of the red list rule.

Detecting the watermark

H_0: The text sequence is generated with no knowledge of the red list rule.

If the null hypothesis is true, then the number of green list tokens, denoted $|s|_G$, has expected value $\frac{T}{2}$ and variance $\frac{T}{4}$. The z-statistic for this test is $z = \frac{2\left(|s|_G - \frac{T}{2}\right)}{\sqrt{T}}$.

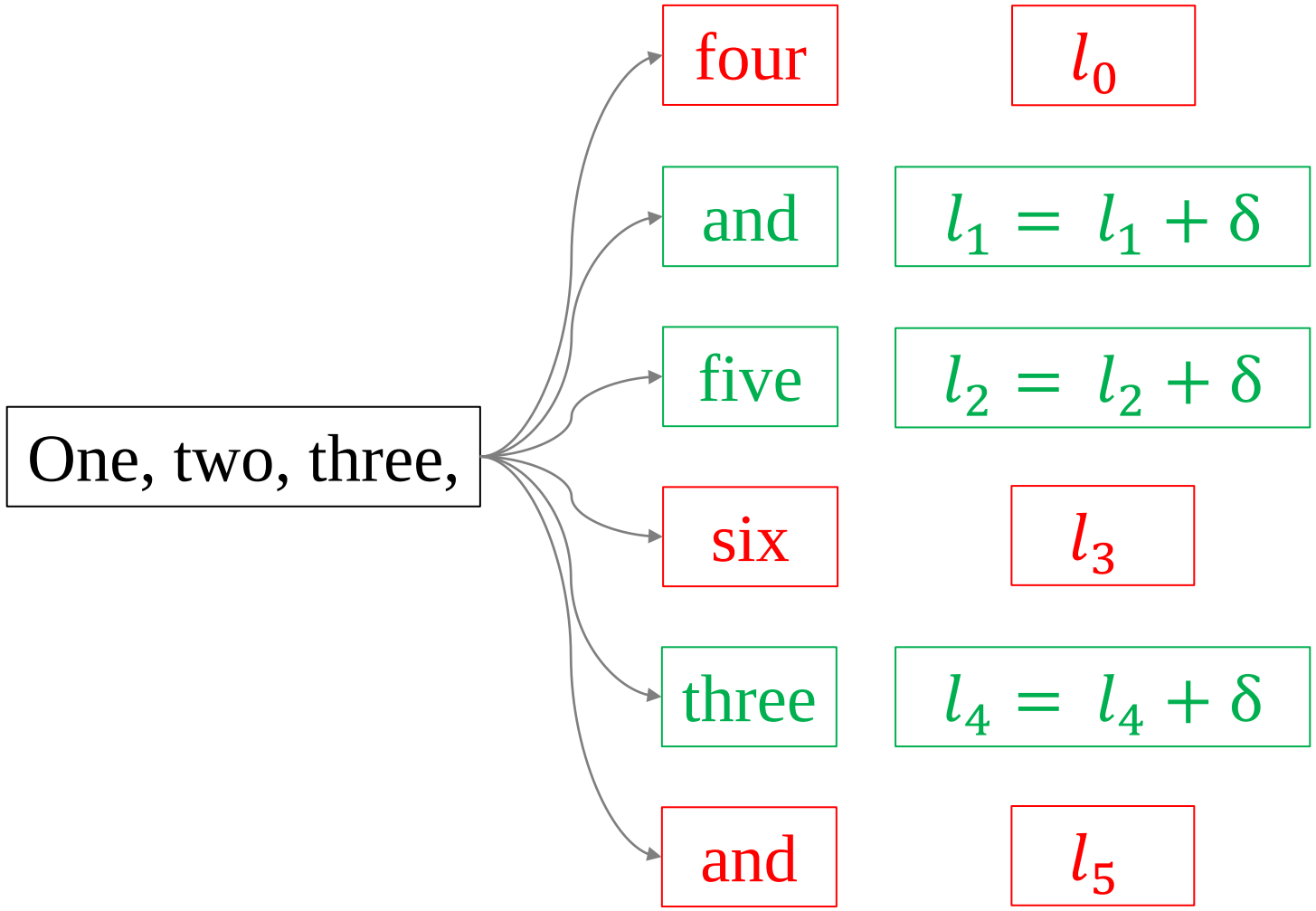
How hard is it to remove the watermark?

If length of a watermarked sequence is T = 1000, amount of modified tokens is n = 200, then z-statistic is

$$z = \frac{2\left(600 - \frac{1000}{2}\right)}{\sqrt{1000}} \approx 6.3$$

and a p-value is $p \approx 10^{-10}$.

«Soft» watermark algorithm



«Soft» watermark algorithm

for t = 0, 1, · · · do

1. Apply the language model to prior tokens $s^{(-Np)} \cdots s^{(t-1)}$ to get a logit vector $l^{(t)}$ over the vocabulary.
2. Compute a hash of token $s^{(t-1)}$, and use it to seed a random number generator.
3. Using this random number generator, randomly partition the vocabulary into a "green list" G of size $\gamma|V|$, and a "red list" R of size $(1 - \gamma)|V|$.
4. Add δ to each green list logit. Apply the softmax operator to these modified logits to get a probability distribution over the vocabulary.

$$\hat{p}_k^{(t)} = \begin{cases} \dfrac{exp(l_k^{(t)} + \delta)}{\sum_{i \in R} exp\left(l_i^{(t)}\right) + \sum_{i \in G} exp\left(l_i^{(t)} + \delta\right)} k \in G \\ \dfrac{exp(l_k^{(t)})}{\sum_{i \in R} exp\left(l_i^{(t)}\right) + \sum_{i \in G} exp\left(l_i^{(t)} + \delta\right)}, k \in R. \end{cases}$$

5. Sample the next token, $s^{(t)}$, using the watermarked distribution $\hat{p}^{(t)}$.

end for

Detecting the watermark

H_0: The text sequence is generated with no knowledge of the red list rule.

For arbitrary γ we have $z = \frac{|s|_G - \gamma T}{\sqrt{T\gamma(1-\gamma)}}$

Analysis of the «soft» watermark

Spike entropy

Given a discrete probability vector p and a scalar z, we define the **spike entropy** of p with modulus z as $S(p,z) = \sum_k \frac{p_k}{1+zp_k}$.

Spike entropy

Given a discrete probability vector p and a scalar z, we define the **spike entropy** of p with modulus z as $S(p,z) = \sum_k \frac{p_k}{1+zp_k}$.

- the spike entropy assumes its minimal value of $\frac{1}{1+z}$, when the entire mass of p is concentrated at a single location

Spike entropy

Given a discrete probability vector p and a scalar z, we define the **spike entropy** of p with modulus z as $S(p,z) = \sum_k \frac{p_k}{1+zp_k}$.

- the spike entropy assumes its minimal value of $\frac{1}{1+z}$, when the entire mass of p is concentrated at a single location
- the spike entropy assumes its maximal value of $\frac{N}{N+z}$, when the mass of p is uniformly distributed

Theorem: Consider watermarked text sequences of T tokens. Each sequence is produced with the algorithm discussed before. Consider there is a raw probability vector $p^{(t)}$, a random green list of size γN, and parameter δ. Define $\alpha = exp(\delta)$ and $|s|_G$ denote the number of green list tokens in sequence s.

Theorem: Consider watermarked text sequences of T tokens. Each sequence is produced with the algorithm discussed before. Consider there is a raw probability vector $p^{(t)}$, a random green list of size γN, and parameter δ. Define $\alpha = exp(\delta)$ and $|s|_G$ denote the number of green list tokens in sequence s.

If a randomly generated watermarked sequence has average spike entropy at least S^*, i.e., $\frac{1}{T}\sum_t S\left(p^{(t)}, \frac{(1-\gamma)(\alpha-1)}{1+(\alpha-1)\gamma}\right) \geq S^*$, then the number of green list tokens in the sequence has expected value at least $E|s|_G \geq \frac{\gamma\alpha T}{1+(\alpha-1)\gamma} S^*$. Furthermore, the number of green list tokens has variance at most $Var|s|_G \leq T\frac{\gamma\alpha S^*}{1+(\alpha-1)\gamma}(1 - \frac{\gamma\alpha S^*}{1+(\alpha-1)\gamma})$.

Table 1. Selected outputs from non-watermarked (NW) and watermarked (W) multinomial sampling using $\gamma = 0.5$ and $\delta = 2.0$. The example in the first row has high entropy and correspondingly high z-scores, without any perceptible degradation in output quality. *The lower row is a failure case where the watermark is too weak to be detected* – it has low entropy and corresponding low z-scores.

prompt	real completion	no watermark (NW)	watermarked (W)	S	(W) z	(NW) PPL	(W) PPL
...tled out of court and publicly reconciled.\nIn the '80s the band's popularity waned in the United States but remained strong abroad. Robin released three solo albums, with limited success. The Bee Gees	returned with some moderate hits in the late 1990s and were inducted into the Rock and Roll Hall of Fame in 1997. With his brothers, Mr. Gibb won six Grammys.\nIn addition to his wife and his brother [...continues]	continued to tour, and Barry became a television producer.\nBut in the early '90s, the Bee Gees' popularity remained high. They scored a hit with "Don't Stop Believing" in 1990, and in 1992 the Bee Ge[...continues]	' 1990 album, "Spirits of the Century," was a mixed critical and commercial success.\nWhen the brothers were nominated for a Grammy Award in 1990, Mr. Gibb's "You Should Be Dancing" and "Massachusetts,[...continues]	0.68	12.73	3.15	1.93
...cond season at Hall Bros Oval.\nThe defender also admitted his surprise at Young's run to the finals but credited the injection of youth into the side.\n"We were really in a building phase last year and	we copped a few floggings with all those juniors blokes coming in," Galvin said.\n"Now, we've kept that core group together for two years and I think we've come along quicker than we anticipated.\nROCK[...continues]	we copped a few floggings with all those juniors blokes coming in," Galvin said.\n"Now, we've kept that core group together for two years and I think we've come along quicker than we anticipated.\n"Tha[...continues]	we copped a few floggings with all those juniors blokes coming in," Galvin said.\n"Now, we've kept that core group together for two years and I think we've come along quicker than we anticipated.\n"Tha[...continues]	0.58	-1.13	1.05	1.04

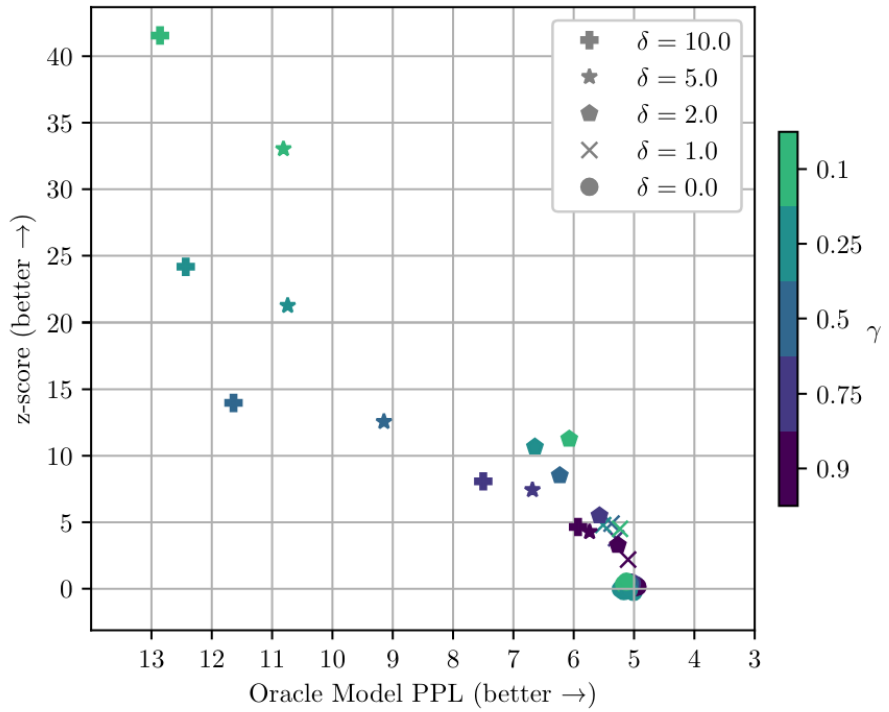
Private Watermarking

A watermark can also be operated in private mode, in which the algorithm uses a random key that is kept secret.

If the attacker has no knowledge of the key used to produce the red list, it becomes more difficult for the attacker to remove the watermark.

Experiments (finally few images!!)

Watermark Strength vs Text Quality



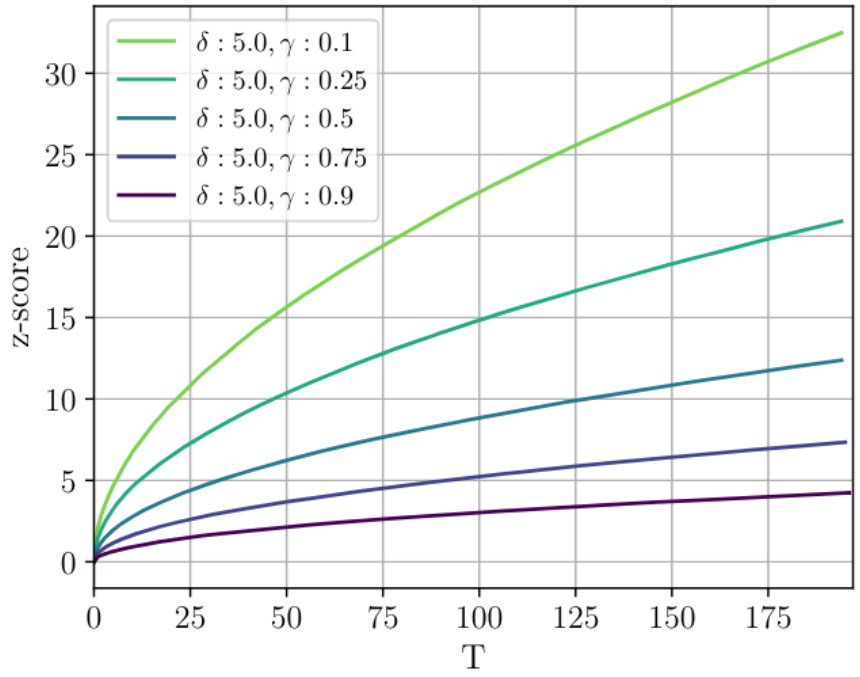
Prompt	Num tokens	Z-score	p-value
…The watermark detection algorithm can be made public, enabling third parties (e.g., social media platforms) to run it themselves, or it can be kept private and run behind an API. We seek a watermark with the following properties:			
No watermark Extremely efficient 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	56	.31	.38
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.	36	7.4	6e-14

$\gamma = 0.1$ is pareto-optimal

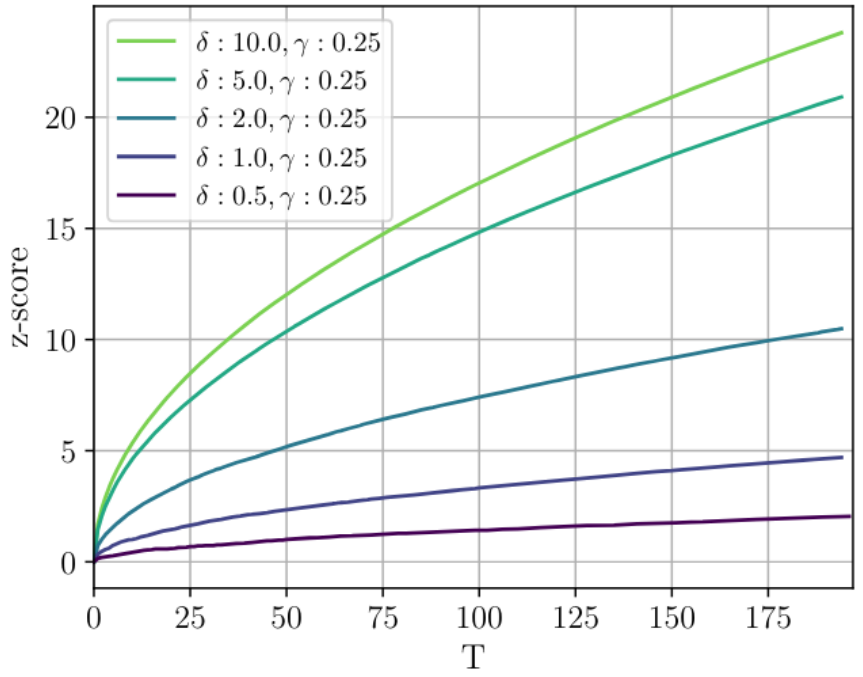
$\gamma, \delta = (0.25, 2)$

Experiments (finally few images!!)

Watermark Strength vs Number of Tokens



(a)



(b)

(a) The dependence of the z-score on the green list size parameter γ

(b) The effect of δ on z-score

Experiments (finally few images!!)

Performance and Sensitivity

z=4.0

δ	γ	FPR	TNR	TPR	FNR
1.0	0.50	0.0	1.0	0.767	0.233
1.0	0.25	0.0	1.0	0.729	0.271
2.0	0.50	0.0	1.0	0.984	0.016
2.0	0.25	0.0	1.0	0.994	0.006
5.0	0.50	0.0	1.0	0.996	0.004
5.0	0.25	0.0	1.0	1.000	0.000

Watermark for language models

Ponomarchuk Anna