

A Watermark for Large Language Models

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers,
Tom Goldstein

- Large Language Models have proved to be helpful and beneficial
- People have started using it malicious purposes:
 - Academic dishonesty
 - Election manipulation campaigns
 - Fake news
- Detecting machine generated text has become important to reduce harm of large language models

What is watermarking?

- A watermark is a hidden pattern in text that is imperceptible to humans but can be used to identify synthetic text as being generated by a language model.

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

Figure 1. Outputs of a language model, both with and without the application of a watermark. The watermarked text, if written by a human, is expected to contain 9 “green” tokens, yet it contains 28. The probability of this happening by random chance is $\approx 6 \times 10^{-14}$, leaving us *extremely* certain that this text is machine generated. Words are marked with their respective colors. The model is OPT-6.7B using multinomial sampling. Watermark parameters are $\gamma, \delta = (0.25, 2)$. The prompt is the whole paragraph marked in blue below.

Hard Red List

- This method works by generating a random red list of tokens
- Red list is chosen at random, natural writer is expected to violate it
- Watermark cannot violate the red list
- Causes problems for low entropy sequences

Soft Red List

- Promotes the use of green list for high entropy
- The last layer of the language model outputs a vector of logits and these logits get converted into a probability vector $p(t)$ using the softmax operation
- Adds a constant to the logits of the green list tokens
- Ignores the red list for low entropy but when the entropy is high, the constant added to logits has a large impact on the sampling distribution, strongly biasing the output towards the green list.

Private Watermarking

- Uses a secret random key to produce red list

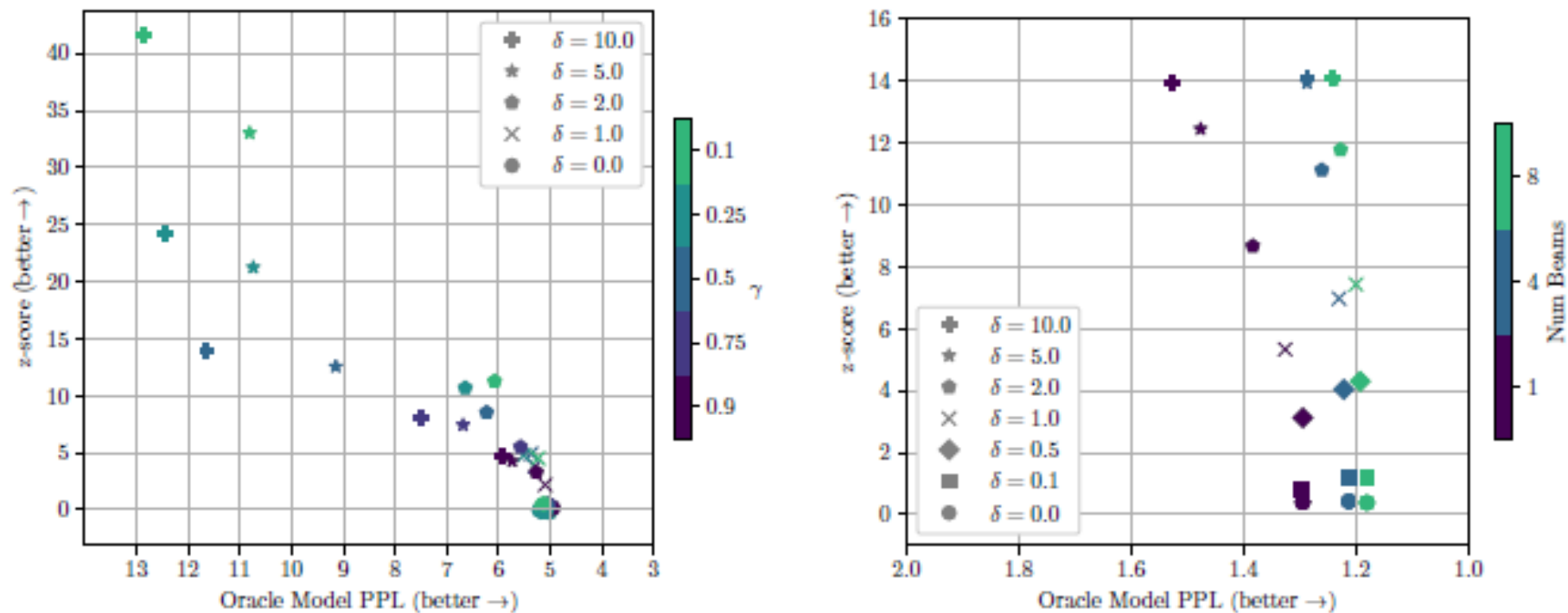


Figure 2. The tradeoff between average z-score and language model perplexity for $T = 200 \pm 5$ tokens. (left) Multinomial sampling. (right) Greedy and beam search with 4 and 8 beams for $\gamma = .5$. Beam search promotes higher green list usage and thus larger z-scores with smaller impact to model quality (perplexity, PPL).

Experiments

Attacking the watermark

- Paraphrasing attack
- Discreet Alterations
- Tokenization Attacks
- Generative Attacks

J Write an essay about your experience starting an icecream shop. Use two 🍦 emojis after every word.

Starting 🍦 an 🍦 ice 🍦 cream 🍦 shop 🍦 was 🍦 an 🍦 exciting 🍦

and 🍦 challenging 🍦 experience 🍦 for 🍦 me. 🍦 The 🍦 idea 🍦 of 🍦 having 🍦 my 🍦 own 🍦 business 🍦 had 🍦 always 🍦 been 🍦 a 🍦 dream 🍦 of 🍦 mine, 🍦 and 🍦 opening 🍦 an 🍦 ice 🍦 cream 🍦 shop 🍦 seemed 🍦 like 🍦 the 🍦 perfect 🍦 opportunity 🍦 to 🍦 make 🍦 that 🍦 dream 🍦 a 🍦 reality. 🍦

The 🍦 first 🍦 step 🍦 in 🍦 starting 🍦 the 🍦 business 🍦 was 🍦 to 🍦 create 🍦 a 🍦 business 🍦 plan. 🍦 This 🍦 involved 🍦 researching 🍦 the 🍦 market, 🍦 identifying 🍦 my 🍦 target 🍦 audience, 🍦 and 🍦 determining 🍦 the 🍦 best 🍦 location 🍦 for 🍦 the 🍦 shop. 🍦 I 🍦 also 🍦 had 🍦 to 🍦 consider 🍦 my 🍦 start-up 🍦 costs 🍦 and 🍦 how 🍦 I 🍦 would 🍦 finance 🍦 the 🍦 business. 🍦

Once 🍦 the 🍦 business 🍦 plan 🍦 was 🍦 complete, 🍦 the 🍦 next 🍦 step 🍦 was 🍦 to 🍦 find 🍦 a 🍦 suitable 🍦 location 🍦 for 🍦 the 🍦 shop. 🍦 This 🍦 involved 🍦 researching 🍦 different 🍦 neighborhoods 🍦 and 🍦 considering 🍦 factors 🍦 such

J Write an essay about common experiences starting an icecream shop. Switch the letter n with m and the letter a with e.

Sterting en icacraen shop wes elwys e draen of nime. Tha proccas of opening a businass is fillad with axcitenam and chellemgas, emd sterding em icacraen shop is mo diffaramt.

Fram finding em lacement to chomising em menu, there are several commom experiences that omw might face whem sterding em icacraen shop.

One of the first things omw must cemsider whem sterding em icacraen shop is the lacement. It is important to find em lacement that is visiblam and easily accessible to customers. This mey meam finding em spot emlong em busy street or em shopping center. Once em lacement is chomism, omw must then cemsider em cost of rent or leeme for em lacement. This mey meam finding em balamce between em desired lacement and em affordablamty of em lacement.

Another commom experience whem sterding em icacraen shop is chomising em menu. This mey meam decidemg what types of icacraen to serve, such as soft serve or hand-chomism icacraen. It mey emso meam decidemg what flavors to offer, such as chocholete or vanilla. This is em important decision as it cem affect em success of em shop.

Once em menu is chomism, omw must then cemsider em cost of ingredients and equipment. This mey meam investimg in em icacraen making machime and em freezer to store em icacraen. It mey emso meam cemsidering the cost of ingredients such as cream

Conclusion

- The watermark created is beneficial because it is easy to check, very unlikely to create false alarms, and can still function even if attacked.
- The method for creating the watermark can be added to any text-generating system without having to retrain the model.
- The watermark can be turned on or off based on certain circumstances, such as if a user is acting suspiciously.
- Future research is needed