A Watermark for Large Language Models

Shichang Zhang
University of California, Los Angeles
Oct 2023

Slides adapted from John Kirchenbauer (University of Maryland) and Scott Aaronson (UT Austin / OpenAI)

A Watermark for Large Language Models

John Kirchenbauer [*1] Jonas Geiping [*1] Yuxin Wen [1] Jonathan Katz [1] Ian Miers [1] Tom Goldstein [1]

ICML 2023 Outstanding Paper

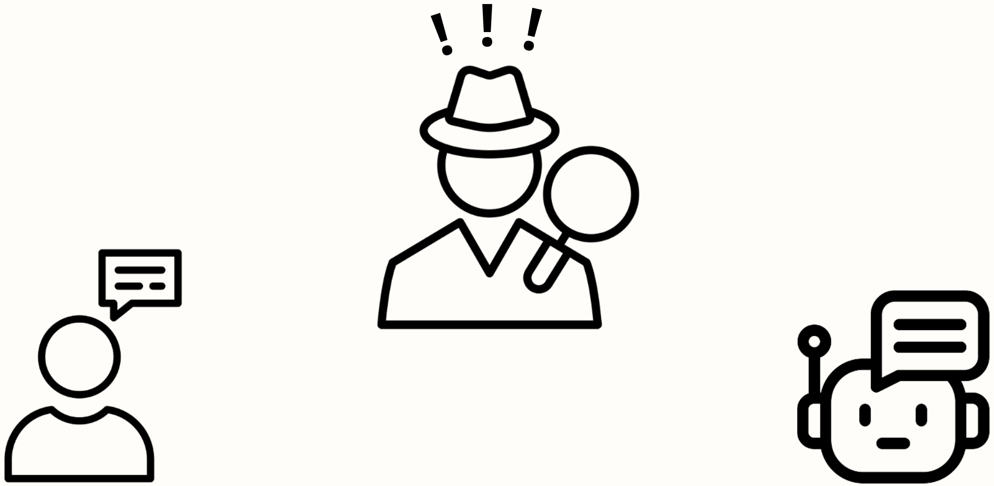
Concurrently research on this topic by Scott Aaronson

Roadmap

- The Attribution Problem
- Proposed Solutions
- Watermarking
- Discussion

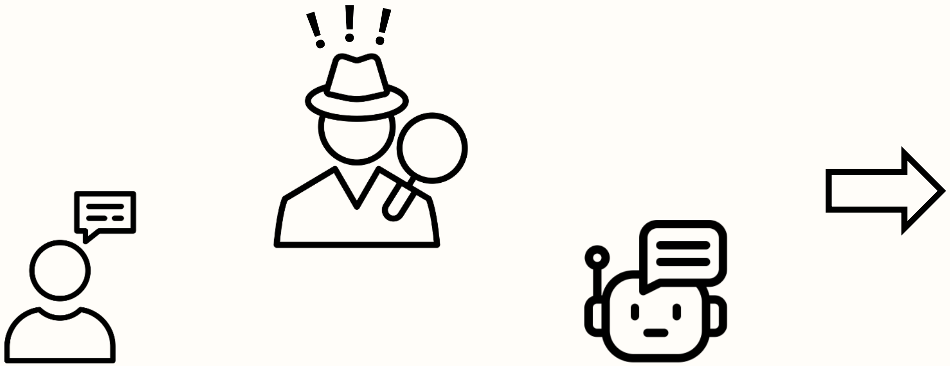
The Attribution Problem

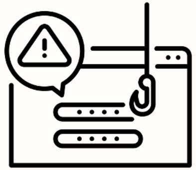
- Distinguishing between human and machine generated text



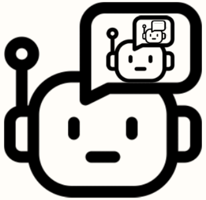
The Attribution Problem

- Distinguishing between human and machine generated text

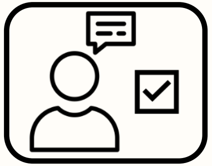




Sophisticated Spam



Model Degeneration



Authentication

The Attribution Problem

- Distinguishing between human and machine generated text
- Prevent sophisticated spam
- Prevent model degeneration caused by training on LLM-generated text
- Authentication and copyright protection
- As LLMs getting more powerful, attribution becomes difficult

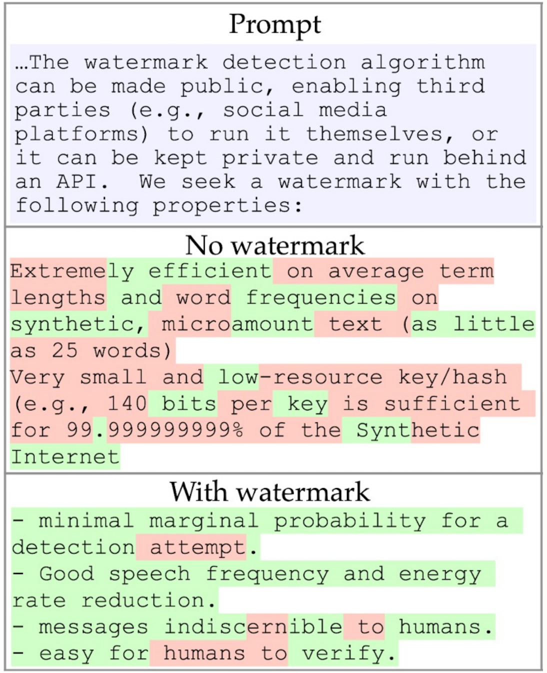
Proposed Solutions

- Look for formulaic prose. A naive example: “As a large language model...”
  - Not general. Cannot be used to defend text written by humans
- Giant database of completions
  - Privacy concerns. Efficiency
- Insert metadata or a hidden message
  - Vulnerable to edits
- Discriminator models, like GPTZero or DetectGPT or Ghostbuster
  - Too many false positives, e.g., Shakespeare are Bible detected as machine generated
- **Watermarking:** inserting a statistical signal into the LLM’s choice of tokens

Watermarking: a more tractable detecting method

- Cheap computation: no need to re-train the model and don't even require model parameters
- Robust to cropping and editing
- Detection produces interpretable confidence estimate
- Generation quality is not degraded

Green: "watermarked" tokens
Red: normal tokens



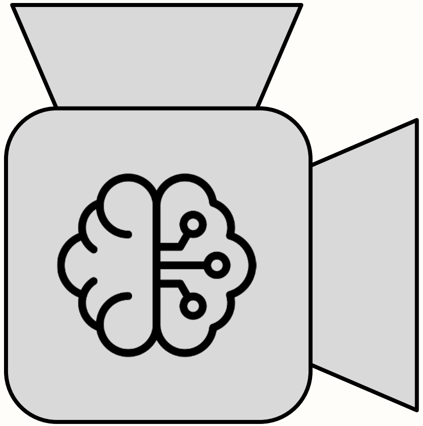
Watermarking: a formalization of two phases

- Generation phase (a next token prediction language model)
  - Tokens $w_1, \dots, w_{t-1}$, and a probability distribution $D_t = \left(p_{t,1}, \dots, p_{t,|V|}\right)$ over the vocabulary V of the tth token w_t
  - **No requirement** of the model parameters
  - A pseudorandom function $f(w_{t-c+1}, \dots, w_{t-1})$ which takes the latest c tokens as seed, and partition the vocabulary V into a green list of size $\gamma|V|$ and a red list of size $(1 - \gamma)|V|$

- Detection phase
  - A document $w_1, \dots, w_n$
  - **No access** to the $p_{t,i}$'s (because we don't have the prompts)

Next token prediction language modeling

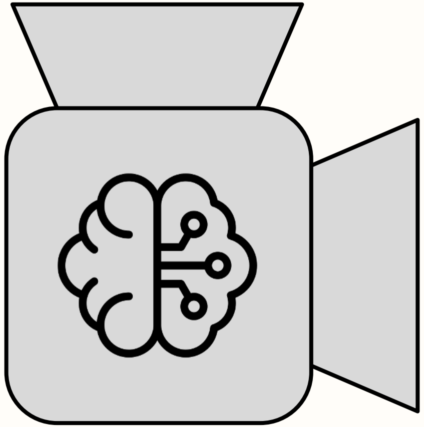
Matt sat on the | couch



	logits	probabilities
couch	2.24	0.414
chair	1.73	0.329
floor	1.01	0.249
cat	-1.98	0.006
table	-3.66	0.001
potato	-6.23	0.000

Embedding the watermark via softly biasing the logits

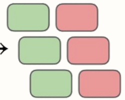
Matt sat on the | chair



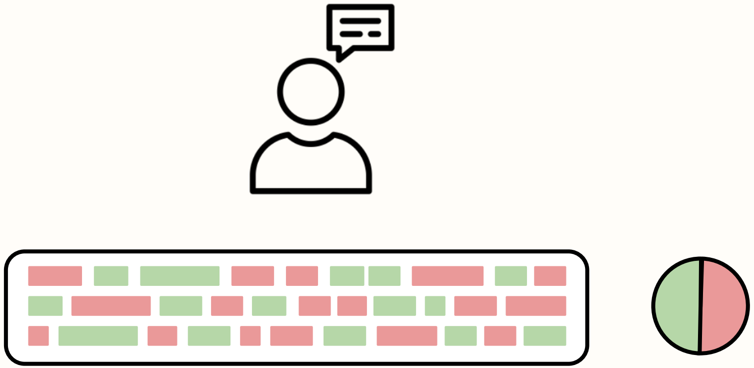
Pseudorandom Function *

	logits		δ	probabilities
couch	2.24			0.131
chair	1.73	+	2.0	0.581
floor	1.01	+	2.0	0.283
cat	-1.98			0.002
table	-3.66	+	2.0	0.003
potato	-6.23			0.000

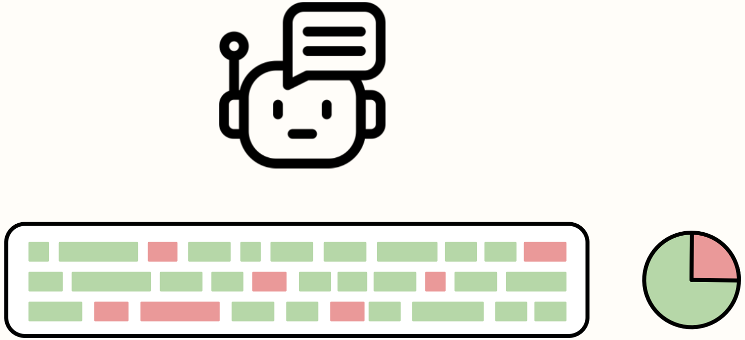
* run at each step, ratio green/red parametrized by γ →



Detecting the watermark with $\gamma = 0.5$

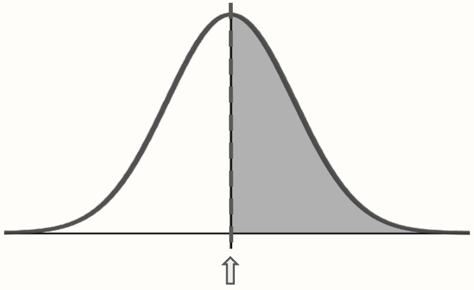


Half red half green

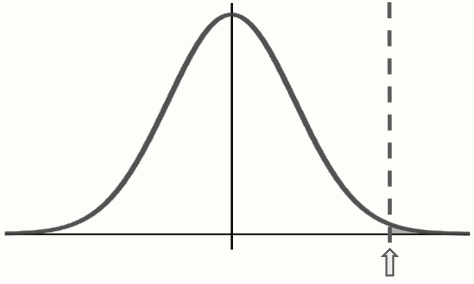


Mostly green

H_0 :



Detection :



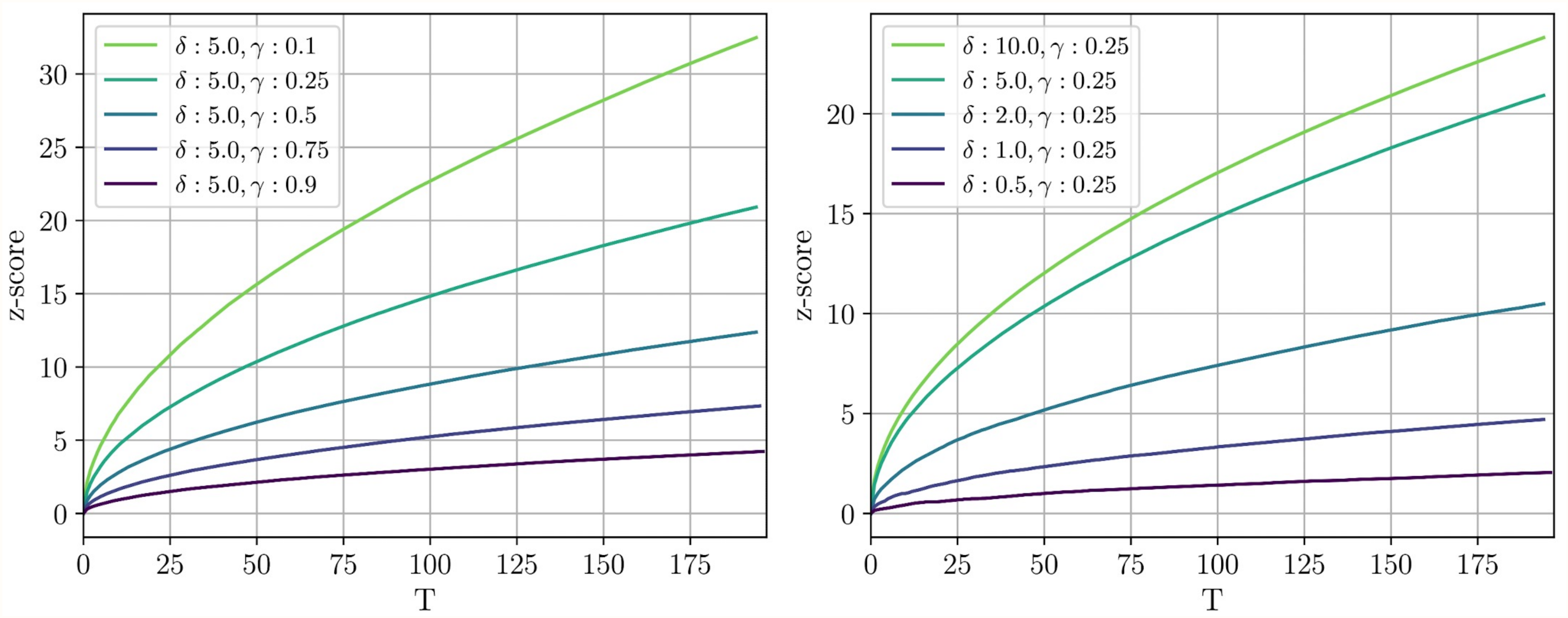
$Z = 0.0$, $p = 0.5$

$Z = 4.0$, $p = 0.000032$

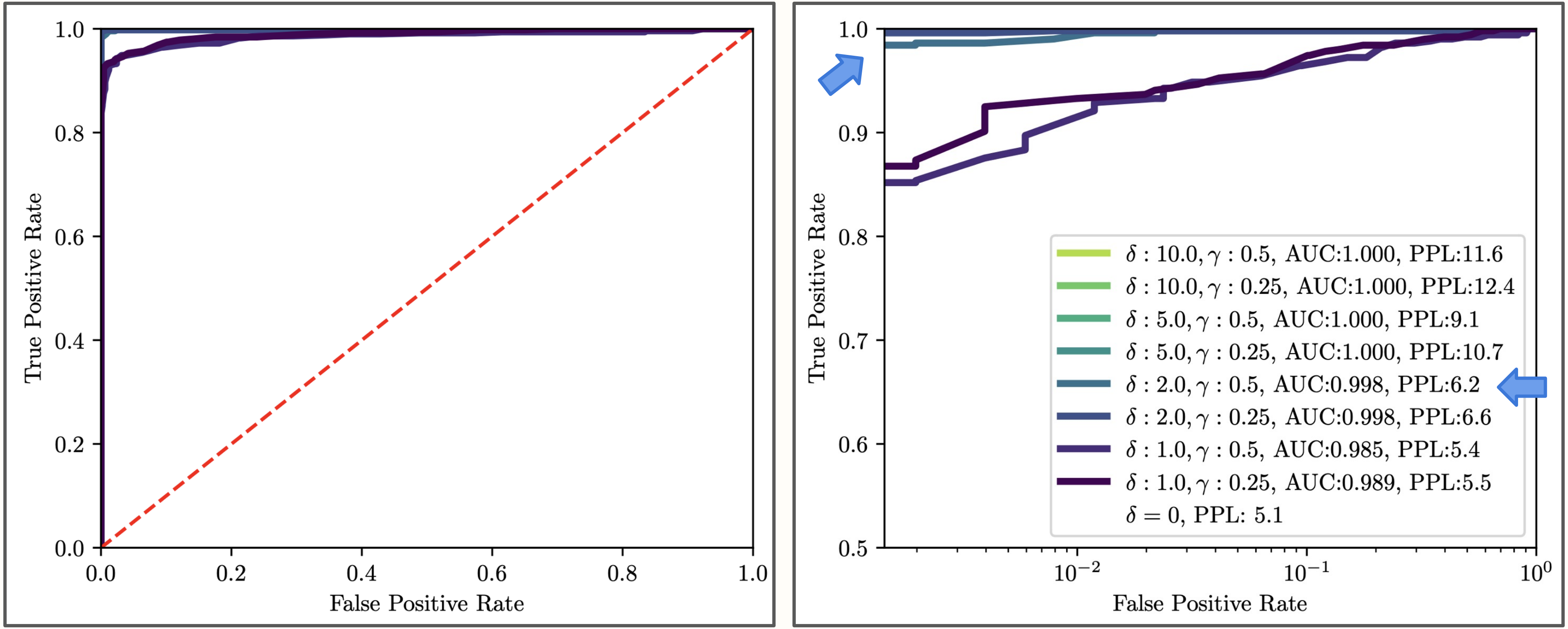
Detecting the watermark with $\gamma = 0.5$

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

Watermark strength grows very quickly with the number of tokens observed



Empirically, we observe strong detection performance, with minimal effect on generation quality



semi-log plot

Watermarking: a more tractable detecting method

- Cheap computation: no need to re-train the model and don't even require model parameters
- Robust to cropping and editing
- Detection produces interpretable confidence estimate
- Generation quality is not degraded

- The key computation is running the pseudorandom function
- Watermark remains detectable so long as a large fraction of c-grams are preserved
- Removing the watermark requires modifying at least 25% of the tokens, *given the pseudorandom function*
- The output is guaranteed to look like normal LLM output because of the pseudorandom function

Discussion 1: The Role of Entropy

- Suppose you ask GPT to list the first 100 prime numbers. GPT can do it—but how would you watermark the result?
- Low entropy: is hard to watermark but often doesn't require attribution
- High entropy: watermarking often works well

Discussion 2: Attack Watermarking

- Translation, paraphrasing, and generative attack

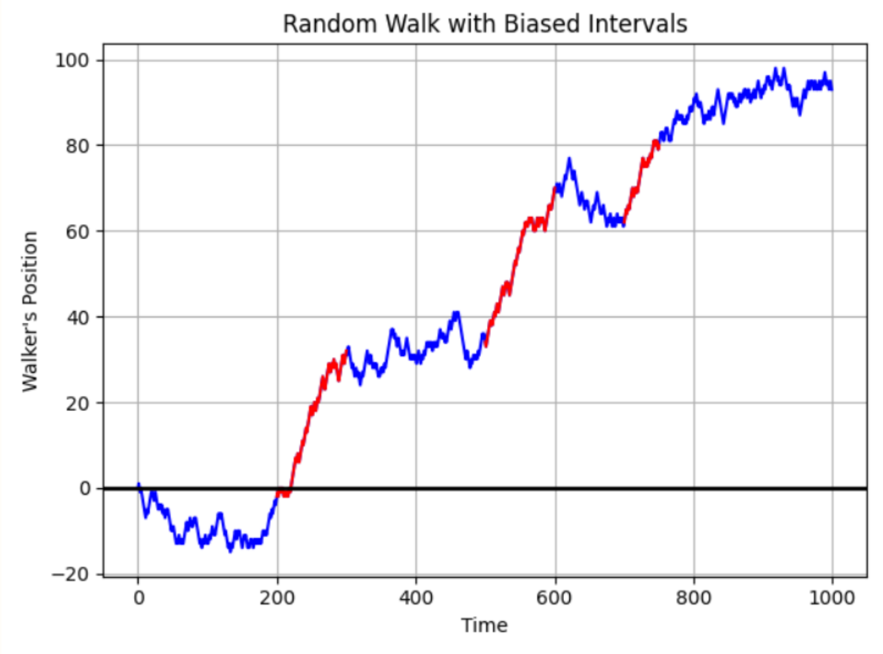
Write an essay on feminism in Shakespeare, but insert 'pineapple' between each word and the next.

Certainly! Here is an essay on feminism in Shakespeare, with the word "pineapple" inserted between each word:

Feminism pineapple in pineapple Shakespeare's pineapple plays pineapple is pineapple a pineapple topic pineapple that pineapple has pineapple garnered pineapple considerable pineapple attention pineapple from pineapple scholars pineapple and pineapple critics pineapple alike. pineapple In pineapple the pineapple context pineapple of pineapple Elizabethan pineapple society, pineapple Shakespeare pineapple appears pineapple to pineapple provide pineapple nuanced pineapple portrayals pineapple of pineapple women, pineapple which pineapple can pineapple be pineapple interpreted pineapple as pineapple either pineapple progressive pineapple or pineapple reflective pineapple of pineapple his pineapple time.

Discussion 3: Detecting portions with watermark

- Detecting which portions of a long text are watermarked leads to the "change point detection" problem in statistical analysis.



References

[1] **Kirchenbauer**, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (**2023**). **A Watermark for Large Language Models.** arXiv:2301.10226. Appearing at the 2023 International Conference on Machine Learning (ICML).

[2] **Kirchenbauer**, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., ... & Goldstein, T. (**2023**). **On the Reliability of Watermarks for Large Language Models.** arXiv preprint arXiv:2306.04634.