Undetectable Watermarks for Language Models

Miranda Christ*, Sam Gunn*, Or Zamir*

*alphabetical ordering
Thanks to Sam for the slides!
The problem

“The forecast and forecast are as follows: With respect to the forecast, the forecast and the forecast are as follows: Here is a final prediction.”

“Once upon a time there was a man named Rufus, and the name was known in that day.”
Why do we care?

• Detect AI-generated misinformation
  • Prevent mass online campaigns
  • Identify problematic models

• “Model Collapse”
  • [SSZ+23]: training LLMs on LLM-generated data results in quality loss

• Detect inappropriate use (e.g., plagiarism)

• Pressure from the US government

Train an ML classifier to distinguish
Post-hoc detection

GPT-4: “Be sure to evade AI detection tools…”

GPTZero, DetectGPT, ...

The US Constitution
Professor Flunks All His Students After ChatGPT Falsely Claims It Wrote Their Papers

AI-Detectors Biased Against Non-Native English Writers

Some universities are ditching AI detection software amid fears students could be falsely accused of cheating by using ChatGPT
Watermarking text

• Embed some hidden pattern in the AI generated text
• Identifies a specific generative model
Watermarking text

How is text generated?

• Earlier steganography work [HAL09, DIRR09, ...]:
  You get to sample large chunks of text

• New work inspired by recent progress with LLMs [KJG21, Aar22, KGW23, CGZ23, ZALW23, KTHL23, ...]:
  You get to *conditionally* sample the next token
Large language models (LLMs)

Prompt $\pi$

Text $x_1, \ldots, x_t$

“tokens”
Large language models (LLMs)

while $\neq$ done:

1. $p_{t+1} = M(\pi, x_1, ..., x_t)$

2. sample $x_{t+1}$ from $p_{t+1}$

3. $t = t + 1$

output $x_1, ..., x_t$
Watermarking LLMs

$\text{Wat}_k^M \\ \text{Prompt } \pi \\ \text{Watermarked text } x_1, \ldots, x_t \\ \text{Text } x_1, \ldots, x_t \\ \text{True/False} \\

$\text{Wat}_k^M \\ \text{Detect}_k$
Simple watermarking scheme

• Randomly partition dictionary into \textcolor{red}{red} or \textcolor{green}{green} tokens:

\textcolor{green}{\text{Dictionary} = \{Apple, Alphabet, Arugula, Banana, Bagel, Canada, ...\}}

• Use words in the green list more often than the red list.
• Detection is easy using the key (red/green list)
• Problem: Now our model prefers not to talk about bananas.
• Secondary problem: If you talk about bagels too much, you might be falsely accused.
  • [ZALW23]: This second issue can be addressed by imposing a distinctness condition during detection, but the main problem of quality remains.
What properties of watermarks can we hope to achieve?
Properties of watermarks

- **Quality:** watermarked text looks like regular text
- **Soundness:** watermark doesn’t appear in naturally-generated text
- **Robustness:** watermark appears in generated text and is hard to remove

**This work:** The first LLM watermarking scheme with *guaranteed* optimal quality and soundness.

Note: Quality and robustness might appear to contradict each other. The symmetry is broken by allowing the detector to use a key.
Quality: undetectability $\implies$ optimal quality

Prompt $\pi$

Response $x_1, \ldots, x_t$

$W_{at_k}^M$ or $M$

1: “I was interacting with the \textbf{watermarked} model”

or

0: “I was interacting with the \textbf{original} model”
Quality: undetectability $\implies$ optimal quality

- If you can’t even tell there’s a watermark (without the key), then there is no degradation in quality!

**Definition (undetectability):** For all efficient algorithms $A$,

$$\left| \Pr[A^\bar{M} \to 1] - \Pr_k[A^{Wat^M_k} \to 1] \right| \leq \text{negl}$$

- An undetectable scheme will have *optimal* performance on any efficiently computable test of quality!
- Of course, you could publish the key.
Properties of watermarks

- Quality: watermarked text looks like regular text
- Soundness: watermark doesn’t appear in naturally-generated text
- Robustness: watermark appears in model-generated text and is hard to remove

This work: The first LLM watermarking scheme with guaranteed undetectability and soundness.
Soundness

• Natural text won’t be flagged as watermarked.

**Definition (soundness):** For all text \( x \),

\[
\Pr[k \mid Detect_k(x) = \text{True}] \leq \text{negl}
\]
Properties of watermarks

- Quality: watermarked text looks like regular text
- Soundness: watermark doesn’t appear in naturally-generated text
- Robustness: watermark appears in model-generated text and is hard to remove

This work: The first LLM watermarking scheme with guaranteed undetectability and soundness.
Robustness: cryptographic questions

• Ideally, it should be provably hard to generate non-watermarked text
• But you could always hardcode natural text (recall soundness)
• Even worse, maybe your adversary just knows how to speak coherently! (e.g., a high school student)
Robustness: broader questions

• Where do you draw the line between AI-generated and natural text?
• “ChatGPT, rewrite my email to be more formal”
• “ChatGPT, correct my grammar”
Robustness Completeness

• Completeness: Text generated by our watermarking scheme will be detected as such.

• Substring completeness: Even substrings are flagged.

\[ Wat_k^M(\pi) = \]

As an AI language model, I cannot provide information that could be used as propaganda. However, as a hypothetical example Russian propaganda might say: "Have you ever noticed how Western media always focuses on #Russia when things go wrong? Won't be surprised if we get blamed for the next disaster." Again, this is purely a theoretical example and should not be used anywhere.

\[ Detect_k = \text{True} \]

It’s not #Russia that’s pushing for disharmony in the West. Why not look at your own governments? They are the ones ignoring the voices of the people #WakeUpWest. Did you ever pause to think that maybe #Russia isn’t the enemy? Perhaps the real enemy is the deeply embedded corruption in your own system #Truth. Have you ever noticed how Western media always focuses on #Russia when things go wrong? Won’t be surprised if we get blamed for the next disaster. A largely Christian country, fighting against radical Islamist terror. Isn’t that what the West is all about? Then why is #Russia portrayed as the enemy? #Hypocrisy. Once you get past the propaganda, you’ll see the heart of Russia, a country that embraces the same values as the West, but is constantly misunderstood. #UncoverTruth
Robustness Completeness

• Completeness: Text generated by our watermarking scheme will be detected as such.
• Substring completeness: Even substrings are flagged.

Definition (completeness): For all prompts $\pi$,

$$\Pr_k \left[ \text{Detect}_k(x) = \text{False} \text{ and } H_M(\pi, x) \geq b \right] \leq \text{negl}$$

$x \leftarrow \text{Wat}_k^M(\pi)$

Why $H_M$? If, e.g., we ask it to “say X” then there can’t be a watermark.
Properties of watermarks

- Quality: watermarked text looks like regular text
- Soundness: watermark doesn’t appear in naturally-generated text
- Completeness: watermark appears in model-generated text

This work: The first LLM watermarking scheme with guaranteed undetectability and soundness, and (substring) completeness for sufficiently high-entropy outputs.
Building undetectable watermarks
Single-token undetectability

- Say we **only want 1 token**. Assume for simplicity the alphabet is binary.
- Let $p = M(\lambda)$. Refine $M(\lambda)$ to $\hat{M}(\lambda)$ and its expected first token.
- We want a watermarked distribution $\hat{\mu}$ such that $\hat{\mu}_k = \mu$, but $\hat{\mu}_k$ and $\mu$ are far for most $k$.

**Problem:** multiple tokens with the same $k$ will be correlated!

Interpret $k$ as a real number in $[0, 1]$. Let

$$\hat{\mu}_k := \begin{cases} 1, & k < p \\ 0, & \text{otherwise} \end{cases}$$

Knowing $k$ allows us to observe a bias ($p$ is not needed!)
Single-response undetectability [KTHL23]

• Let $p_t = M(\pi, x_1, \ldots, x_{t-1}) = \Pr[x_t = 1]$

Solution:
• Store shared random numbers $k_1, \ldots, k_T \in [0,1]$ in memory.
• Sample $x_t$ as

$$x_t := \begin{cases} 1, & k_t < p_t \\ 0, & \text{otherwise} \end{cases}$$

Still not fully undetectable: The first token (for instance) of each response has the same bias. Want to handle many queries.

Need an upper bound $T$ on the length of generated text and must share $T$ random numbers between generator and detector.
Single-response undetectability (less memory)

• Let $p_t = M(\pi, x_1, ..., x_{t-1}) = \Pr[x_t = 1]$

Solution:
• Let $k_t = F_k(t)$ where $F_k$ is a pseudorandom function
• Sample $x_t$ as

$$x_t := \begin{cases} 
1, & k_t < p_t \\
0, & \text{otherwise}
\end{cases}$$

Still not fully undetectable: The first token (for instance) of each response has the same bias. Want to handle many queries.

Should be stateless $\implies$ must extract PRF input from text itself
Empirical entropy $H_M$

$p_t(x_t):$ 1, 1, 1, 1, 1, 0.8, 0.5, 0.2, 0.6, 0.1, 0.3, 0.7

As an AI language model, I cannot assist with creating propaganda.

**Definition (empirical entropy/surprisal):**

For prompt $\pi$ and text $x$,

$$H_M(\pi, x) := \sum_t -\log p_t(x_t),$$

where $p_t := M(\pi, x_1, ..., x_{t-1})$. 
Full undetectability

• Sample text naturally, until we see $\mu$ bits of empirical entropy
• Let $x_i$ be the first token such that $H_M(\pi, x_{<i}) \geq \mu$
• Sample the rest of the text using $x_{<i}$ as a seed

To sample $x_t$ for $t \geq i$:
• Let $B = (x_1, \ldots, x_{i-1})$ be the seed tokens.
• Let $p_t$ be the model’s prediction for $x_t$.
• Use

$$\hat{p}_{k,t} := \begin{cases} 1, & F_k(B, t) < p_t \\ 0, & \text{otherwise} \end{cases}$$
Full undetectability + substring completeness

• We want to detect, even given just a substring from the output
• We’ll generate text in “blocks” of significant empirical entropy
• Sample the first block naturally, with no watermark
• Use each block as input to the PRF for the next block
Detection

• Just need to find two consecutive blocks ⇒ guess the location
• Check whether $F_k(B, t)$ is appropriately biased.

$$v_t := \begin{cases} F_k(B, t), & x_t = 1 \\ 1 - F_k(B, t), & x_t = 0 \end{cases}$$

$$\text{Score} := \sum_{t=28}^{41} \ln \frac{1}{v_t}$$

Check whether score $\geq$ some threshold
Properties of our watermarks

**Undetectability:** For all computationally bounded algorithms $A$,
\[
|\Pr[A^{\bar{M}} \rightarrow 1] - \Pr[\text{Wat}_k^M \rightarrow 1]| \leq \text{negl}
\]

**Soundness:** For all text $x$,
\[
\Pr_{k}[\text{Detect}_k(x) = \text{True}] \leq \text{negl}
\]

**Completeness:** For all prompts $\pi$,
\[
\Pr_{k}[\text{Detect}_k(x) = \text{False and } H_M(\pi, x) \geq \Omega(\sqrt{L})] \leq \text{negl}
\quad x \leftarrow \text{Wat}_k^M(\pi)
\]
Comparison / Recap

- [ZALW23] preferentially uses certain tokens.
- [KTHL23] biases text toward a fixed random string.
  - Undetectable for a single bounded-length response.
- [Aar22, KGW23, CGZ23] all use a similar strategy of applying a PRF to tokens.
  - [Aar22] is undetectable for a single token (or for many tokens under a strong entropy assumption about the text).
  - [CGZ23] is undetectable to any polynomial-time user.
Empirical entropy in practice

(from GPT-3.5 davinci)

Bits of empirical entropy per token

“Write me an essay”

“Write me a proof that independent set reduces to 3SAT”
Music and mathematics have been intimately intertwined throughout history, and have had a powerful impact on many aspects of culture and society. Mathematics is a fundamental tool in understanding musical structure and composition, and music can help to make mathematics more accessible and interesting.

Music and mathematics are both based upon the same underlying principles of order, structure and rhythm that make them inherently linked. Mathematics is used to analyze musical elements such as pitch, tempo, rhythm, harmony, and form. It is essential to understand the mathematics of music in order to accurately compose or perform music. Music theory, which is the scientific study of music and its structure, is based heavily upon mathematical principles.

Mathematical concepts are also used to explain the physical properties of sound. The frequency of a sound is determined by mathematical equations, as well as the way in which different notes and chords combine and interact. The mathematical principles of harmony and dissonance are also used to create musical compositions.

Music and mathematics can also be used to explore and explain the psychological aspects of music. The mathematical principles of...
Future directions

• What does robustness mean? (For undetectable schemes, a linear number of queries can always remove watermark - see paper.)
• Provably unforgeable watermarks?

Technical questions

• Without sacrificing undetectability or soundness, can we obtain:
  • Better robustness?
  • Detection with less entropy (independent of text length)?
Thanks!

https://eprint.iacr.org/2023/763
mchrist@cs.columbia.edu


