A Statistical Framework of Watermarks for Large Language Models

Weijie Su

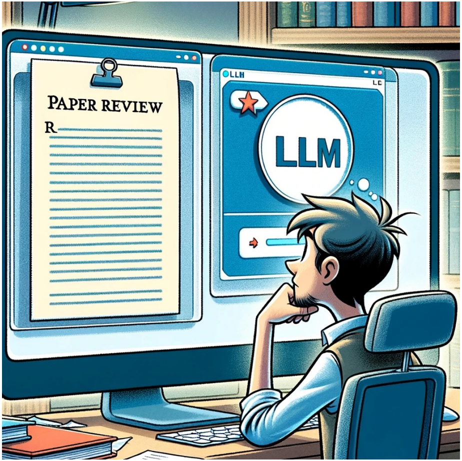
University of Pennsylvania

Do you trust the students?

Did the student complete the homework independently,
or did an LLM assist?

Peer review or LLM-assisted review?

- Liang et al. (2024): 6.5% to 16.9% of some ML conference reviews substantially modified by LLMs
- Is the review genuinely authored by the reviewer or significantly contributed by an LLM?

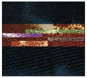


An emerging academic integrity issue

ELSEVIER

Surfaces and Interfaces

journal homepage: www.sciencedirect.com/journal/surfaces-and-interfaces



The three-dimensional porous mesh structure of Cu-based metal-organic-framework - aramid cellulose separator enhances the electrochemical performance of lithium metal anode batteries

[redacted]

A R T I C L E I N F O

Keywords:
Lithium metal battery
Lithium dendrites
CuMOF-ANFs separator

A B S T R A C T

Lithium metal, due to its advantages of high theoretical capacity, low density and low electrochemical reaction potential, is used as a negative electrode material for batteries and brings great potential for the next generation of energy storage systems. However, the production of lithium metal dendrites makes the battery life low and poor safety, so lithium dendrites have been the biggest problem of lithium metal batteries. This study shows that the larger specific surface area and more pore structure of Cu-based metal-organic-framework - aramid cellulose (CuMOF-ANFs) composite separator can help to inhibit the formation of lithium dendrites. After 110 cycles at 1 mA/cm^2, the discharge capacity retention rate of the Li-Cu battery using the CuMOF-ANFs separator is about 96 %. Li-Li batteries can continue to maintain low hysteresis for 2000 h at the same current density. The results show that CuMOF-ANFs composite membrane can inhibit the generation of lithium dendrites and improve the cycle stability and cycle life of the battery. The three-dimensional (3D) porous mesh structure of CuMOF-ANFs separator provides a new perspective for the practical application of lithium metal battery.

1. Introduction

Certainly, here is a possible introduction for your topic:Lithium-metal batteries are promising candidates for high-energy-density rechargeable batteries due to their low electrode potentials and high

chemical stability of the separator is equally important as it ensures that the separator remains intact and does not react or degrade in the presence of the electrolyte or other battery components. A chemically stable separator helps to prevent the formation of reactive species that can further promote dendrite growth. Researchers are actively exploring

It's important to detect LLM-generated text, but how?

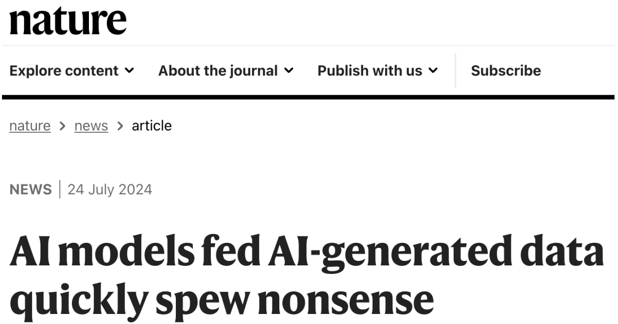
Applications

- Fostering original work in education and maintaining academic integrity

It's important to detect LLM-generated text, but how?

Applications

- Fostering original work in education and maintaining academic integrity
- Preserving the quality of data for training future AI models



nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature > news > article

NEWS | 24 July 2024

AI models fed AI-generated data quickly spew nonsense

It's important to detect LLM-generated text, but how?

Applications

- Fostering original work in education and maintaining academic integrity
- Preserving the quality of data for training future AI models
- Preventing fraud and deception

It's important to detect LLM-generated text, but how?

Applications

- Fostering original work in education and maintaining academic integrity
- Preserving the quality of data for training future AI models
- Preventing fraud and deception

It's important to detect LLM-generated text, but how?

Applications

- Fostering original work in education and maintaining academic integrity
- Preserving the quality of data for training future AI models
- Preventing fraud and deception

- Ad hoc methods leverage context, linguistic patterns, and other markers:
  - Classifiers using synthetic and human text data (GPTZero, 2023; ZeroGPT, 2023)
  - Log probability curvature (Mitchell et al., 2023; Bao et al., 2023)
  - Divergent n-gram analysis (Yang et al., 2023)

It's important to detect LLM-generated text, but how?

Applications

- Fostering original work in education and maintaining academic integrity
- Preserving the quality of data for training future AI models
- Preventing fraud and deception

- Ad hoc methods leverage context, linguistic patterns, and other markers:
  - Classifiers using synthetic and human text data (GPTZero, 2023; ZeroGPT, 2023)
  - Log probability curvature (Mitchell et al., 2023; Bao et al., 2023)
  - Divergent n-gram analysis (Yang et al., 2023)
- Inaccurate, unreliable (Weber-Wulff et al., 2023), and often biased (Krishna et al., 2024; Sadasivan et al., 2023; Liang et al., 2023)

It's important to detect LLM-generated text, but how?

Applications

- Fostering original work in education and maintaining academic integrity
- Preserving the quality of data for training future AI models
- Preventing fraud and deception

- Ad hoc methods leverage context, linguistic patterns, and other markers:
  - Classifiers using synthetic and human text data (GPTZero, 2023; ZeroGPT, 2023)
  - Log probability curvature (Mitchell et al., 2023; Bao et al., 2023)
  - Divergent n-gram analysis (Yang et al., 2023)
- Inaccurate, unreliable (Weber-Wulff et al., 2023), and often biased (Krishna et al., 2024; Sadasivan et al., 2023; Liang et al., 2023)
- LLM-generated text increasingly resembles human-written text!

It seems hopeless...

- Fundamentally impossible to distinguish between LLM-generated and human-written text (based solely on text alone)

A principled approach: watermarking LLM

Hope: LLMs are probabilistic machines, and we *control* how they generate text

A principled approach: watermarking LLM

Hope: LLMs are probabilistic machines, and we *control* how they generate text

A watermark embeds subtle statistical signals into LLM-generated text (Kirchenbauer et al., 2023a)

- Dependence between observed text and certain hidden information for generating text
- Unlikely to appear in human-written text

A (very) active research area with practical importance

A Zoo of Watermarking Schemes (since January 2023):

A (very) active research area with practical importance

A Zoo of Watermarking Schemes (since January 2023):

Kirchenbauer et al. (2023a); Aaronson (2023); Kuditipudi et al. (2023); Zhao et al. (2024b); Fernandez et al. (2023); Christ et al. (2023); Wu et al. (2023); Hu et al. (2023); Kirchenbauer et al. (2023b); Zhao et al. (2024a)

A (very) active research area with practical importance

A Zoo of Watermarking Schemes (since January 2023):
Kirchenbauer et al. (2023a); Aaronson (2023); Kuditipudi et al. (2023); Zhao et al. (2024b); Fernandez et al. (2023); Christ et al. (2023); Wu et al. (2023); Hu et al. (2023); Kirchenbauer et al. (2023b); Zhao et al. (2024a)

- Biden AI executive order

A (very) active research area with practical importance

A Zoo of Watermarking Schemes (since January 2023):
Kirchenbauer et al. (2023a); Aaronson (2023); Kuditipudi et al. (2023); Zhao et al. (2024b); Fernandez et al. (2023); Christ et al. (2023); Wu et al. (2023); Hu et al. (2023); Kirchenbauer et al. (2023b); Zhao et al. (2024a)

- Biden AI executive order
- OpenAI, Google, Meta, and other tech giants have pledged to watermark AI content

Statistical challenges/opportunities in watermark research

Control/estimation of errors

- False positive rate: mistakenly detecting human-written text as LLM-generated
- False negative rate: incorrectly classifying LLM-generated text as human-written

Statistical challenges/opportunities in watermark research

Control/estimation of errors

- False positive rate: mistakenly detecting human-written text as LLM-generated
- False negative rate: incorrectly classifying LLM-generated text as human-written

Evaluation of watermarks

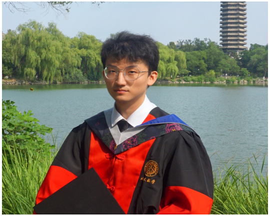
- Comparing different watermarking schemes
- Finding more powerful detection rules
- Robust watermark detection

Team

- *A Statistical Framework of Watermarks for Large Language Models: Pivot, Detection Efficiency and Optimal Rules*. The Annals of Statistics, 2025
- *Robust Detection of Watermarks for Large Language Models Under Human Edits*. arXiv:2411.13868

Team

- *A Statistical Framework of Watermarks for Large Language Models: Pivot, Detection Efficiency and Optimal Rules*. The Annals of Statistics, 2025
- *Robust Detection of Watermarks for Large Language Models Under Human Edits*. arXiv:2411.13868



Xiang Li (Penn)

Feng Ruan (NWU)

Huiyuan Wang (Penn)

Qi Long (Penn)

Outline

1. Preliminaries

2. Hypothesis testing formulation

3. Efficiency and optimal detection

4. Application to Gumbel-max watermark

5. Robust detection

Tokenization

- Tokenization breaks down text into smaller units called “tokens”
- Tokens can be words, parts of words, or even punctuation marks

Tokenization

- Tokenization breaks down text into smaller units called “tokens”
- Tokens can be words, parts of words, or even punctuation marks

Tokens	Characters
122	674

```
The University of Waterloo is a leading public research institution in
 Ontario, Canada, renowned for its strengths in STEM fields, cooperative
 education, and entrepreneurship. Established in 1957, the university is
 home to the world’s largest co-op (work-integrated learning) program,
 allowing students to gain industry experience with top employers such as
 Google, Microsoft, and Tesla. Waterloo is particularly well known for its
 computer science, engineering, and mathematics programs, with the Cherit
on School of Computer Science and the Institute for Quantum Computing (IQC
) driving cutting-edge research in artificial intelligence, cryptography,
 and quantum computing.
```

Autoregressive generation

- Let $\mathcal{W} = \{1, 2, \ldots, K\}$ be the vocabulary and w a token therein
- Vocabulary size $K = |\mathcal{W}|$ is large and varies for different models
- $K = 50{,}257$ for GPT-2/3.5; $32{,}000$ for LLaMA-7B

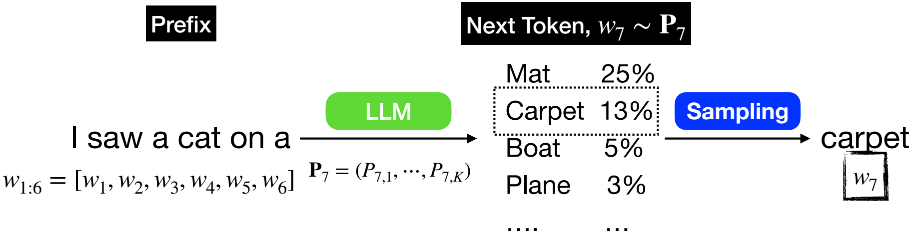
Autoregressive generation

- Let $\mathcal{W} = \{1, 2, \ldots, K\}$ be the vocabulary and w a token therein
- Vocabulary size $K = |\mathcal{W}|$ is large and varies for different models
- $K = 50{,}257$ for GPT-2/3.5; $32{,}000$ for LLaMA-7B
- An LLM generates tokens sequentially by sampling from a (varying) multinomial probability distribution:

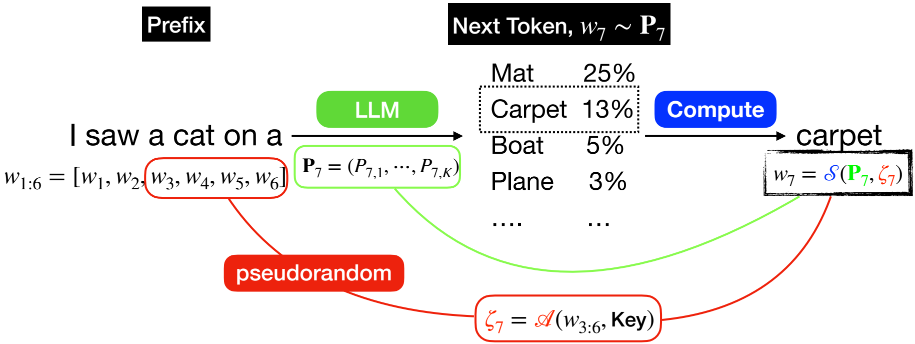
$$w_t \sim \boldsymbol{P}_t$$

  - Next-token prediction (NTP) $\boldsymbol{P}_t = \boldsymbol{P}(w_{1:t-1})$ is a multinomial distribution on $\mathcal{W}$
  - $\boldsymbol{P}_t$ depends also on system prompts, which are unavailable to the public

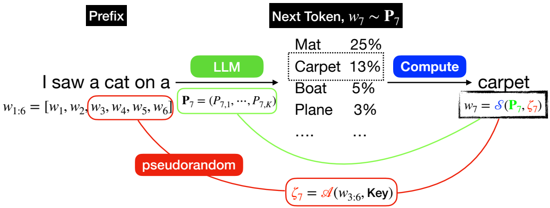
Autoregressive generation: an illustration



Autoregressive generation with watermarks

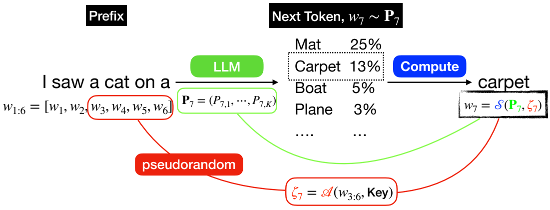


Autoregressive generation with watermarks



- $\mathcal{A}$ is a hash function and $\mathcal{S}(\boldsymbol{P}, \zeta)$ is a (deterministic) decoder

Autoregressive generation with watermarks

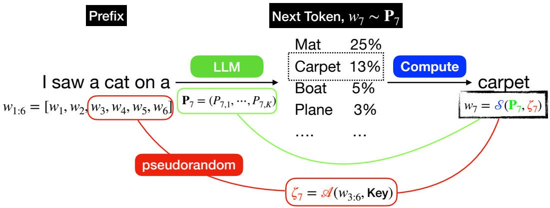


- $\mathcal{A}$ is a hash function and $\mathcal{S}(\boldsymbol{P}, \zeta)$ is a (deterministic) decoder
- Unbiasedness: for any token w,

$$\mathbb{P}(\mathcal{S}(\boldsymbol{P}, \zeta) = w) = P_w$$

Text quality does not degrade

Autoregressive generation with watermarks



- $\mathcal{A}$ is a hash function and $\mathcal{S}(\boldsymbol{P}, \zeta)$ is a (deterministic) decoder
- Unbiasedness: for any token w,

$$\mathbb{P}(\mathcal{S}(\boldsymbol{P}, \zeta) = w) = P_w$$

  Text quality does not degrade
- Watermark is the dependence between w_t and ζ_t!

There is hope

Coupling: the complete observation is

$$(\text{pseudorandomness}, \text{text})$$

and you design the dependence!

A baby watermark

- Let $\mathcal{W} = \{0, 1\}$, $\boldsymbol{P}_t = (P_{t,0}, P_{t,1})$, ζ_t be iid copies of $\mathcal{U}(0, 1)$
- Decoder

$$w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t) = \begin{cases} 0 & \text{if } \zeta_t \leqslant P_{t,0} \\ 1 & \text{if } \zeta_t > P_{t,0} \end{cases}$$

A baby watermark

- Let $\mathcal{W} = \{0, 1\}$, $\boldsymbol{P}_t = (P_{t,0}, P_{t,1})$, ζ_t be iid copies of $\mathcal{U}(0, 1)$
- Decoder

$$w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t) = \begin{cases} 0 & \text{if } \zeta_t \leqslant P_{t,0} \\ 1 & \text{if } \zeta_t > P_{t,0} \end{cases}$$

Unbiasedness

$$\mathbb{P}(\mathcal{S}(\boldsymbol{P}, \zeta) = w) = P_w$$

for $w = 0, 1$

A baby watermark

- Let $\mathcal{W} = \{0, 1\}$, $\boldsymbol{P}_t = (P_{t,0}, P_{t,1})$, ζ_t be iid copies of $\mathcal{U}(0, 1)$
- Decoder

$$w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t) = \begin{cases} 0 & \text{if } \zeta_t \leqslant P_{t,0} \\ 1 & \text{if } \zeta_t > P_{t,0} \end{cases}$$

Unbiasedness

$$\mathbb{P}(\mathcal{S}(\boldsymbol{P}, \zeta) = w) = P_w$$

for $w = 0, 1$

Embedded signal

- If ζ_t is large, w_t is more likely to be 1 instead of 0
- Statistic for detection:

$$\sum_{t=1}^{n} (2w_t - 1)(2\zeta_t - 1)$$

Gumbel-max watermark (Aaronson, 2023)

A watermark corresponds to sampling from a multinomial distribution

Gumbel-max watermark (Aaronson, 2023)

A watermark corresponds to sampling from a multinomial distribution

Gumbel-max trick

Let $\zeta = (U_1, U_2, \ldots, U_K)$ consist of iid copies of $\mathcal{U}(0, 1)$

$$\arg\max_{w \in \mathcal{W}} \frac{\log U_w}{P_w} \sim \boldsymbol{P} \equiv (P_w)_{w \in \mathcal{W}}$$

Gumbel-max watermark (Aaronson, 2023)

A watermark corresponds to sampling from a multinomial distribution

Gumbel-max trick

Let $\zeta = (U_1, U_2, \ldots, U_K)$ consist of iid copies of $\mathcal{U}(0,1)$

$$\arg\max_{w \in \mathcal{W}} \frac{\log U_w}{P_w} \sim \boldsymbol{P} \equiv (P_w)_{w \in \mathcal{W}}$$

Gumbel-max watermark (Aaronson, 2023)

$$\mathcal{S}^{\mathrm{gum}}(\boldsymbol{P}, \zeta) = \arg\max_{w \in \mathcal{W}} \frac{\log U_w}{P_w}$$

- Pseudorandom $\zeta_t = (U_{t,1}, \ldots, U_{t,K}) = \mathcal{A}(w_{1:t-1}, \mathrm{Key})$

Gumbel-max watermark (Aaronson, 2023)

A watermark corresponds to sampling from a multinomial distribution

Gumbel-max trick

Let $\zeta = (U_1, U_2, \ldots, U_K)$ consist of iid copies of $\mathcal{U}(0,1)$

$$\arg\max_{w \in \mathcal{W}} \frac{\log U_w}{P_w} \sim \boldsymbol{P} \equiv (P_w)_{w \in \mathcal{W}}$$

Gumbel-max watermark (Aaronson, 2023)

$$\mathcal{S}^{\mathrm{gum}}(\boldsymbol{P}, \zeta) = \arg\max_{w \in \mathcal{W}} \frac{\log U_w}{P_w}$$

- Pseudorandom $\zeta_t = (U_{t,1}, \ldots, U_{t,K}) = \mathcal{A}(w_{1:t-1}, \mathrm{Key})$
- Embedded signal: selected U_{t,w_t} tends to be larger

Gumbel-max watermark (Aaronson, 2023)

A watermark corresponds to sampling from a multinomial distribution

Gumbel-max trick

Let $\zeta = (U_1, U_2, \ldots, U_K)$ consist of iid copies of $\mathcal{U}(0,1)$

$$\arg\max_{w \in \mathcal{W}} \frac{\log U_w}{P_w} \sim \boldsymbol{P} \equiv (P_w)_{w \in \mathcal{W}}$$

Gumbel-max watermark (Aaronson, 2023)

$$\mathcal{S}^{\text{gum}}(\boldsymbol{P}, \zeta) = \arg\max_{w \in \mathcal{W}} \frac{\log U_w}{P_w}$$

- Pseudorandom $\zeta_t = (U_{t,1}, \ldots, U_{t,K}) = \mathcal{A}(w_{1:t-1}, \text{Key})$
- Embedded signal: selected U_{t,w_t} tends to be larger
- Implemented internally at OpenAI

It’s already behind the scenes…

THE WALL STREET JOURNAL.

Subscribe | Sign In

English Edition ▾ | Print Edition | Video | Audio | Latest Headlines | More ▾

Latest World Business U.S. Politics Economy **Tech** Markets & Finance Opinion Arts Lifestyle Real Estate Personal Finance Health Style Sports

EXCLUSIVE

There’s a Tool to Catch Students Cheating With ChatGPT. OpenAI Hasn’t Released It.

Technology that can detect text written by artificial intelligence with 99.9% certainty has been debated internally for two years

Inverse transform watermark (Kuditipudi et al., 2023)

A watermark corresponds to sampling from a multinomial distribution

Inverse transform watermark (Kuditipudi et al., 2023)

A watermark corresponds to sampling from a multinomial distribution

Probability 101: any univariate distribution can be sampled by applying the inverse CDF to $\mathcal{U}([0, 1])$

Inverse transform watermark (Kuditipudi et al., 2023)

A watermark corresponds to sampling from a multinomial distribution

Probability 101: any univariate distribution can be sampled by applying the inverse CDF to $\mathcal{U}([0,1])$

Inverse transform watermark (Kuditipudi et al., 2023)

Let $F(x;\pi) = \sum_{w' \in \mathcal{W}} P_{w'} \cdot \mathbf{1}_{\{\pi(w') \leqslant x\}}$ be the CDF of π-perturbed $\boldsymbol{P}$. Then

$$F^{-1}(U;\pi) = \min\left\{i : \textstyle\sum_{w' \in \mathcal{W}} P_{w'} \cdot \mathbf{1}_{\{\pi(w') \leqslant i\}} \geqslant U\right\}$$

with $U \sim \mathcal{U}(0,1)$ satisfies $\pi^{-1}(F^{-1}(U;\pi)) \sim \boldsymbol{P} \equiv (P_w)_{w \in \mathcal{W}}$

$$\mathcal{S}^{\text{inv}}(\boldsymbol{P},\zeta) := \pi^{-1}(F^{-1}(U;\pi)) \ \text{ where } \ \zeta = (U,\pi)$$

Inverse transform watermark (Kuditipudi et al., 2023)

A watermark corresponds to sampling from a multinomial distribution

Probability 101: any univariate distribution can be sampled by applying the inverse CDF to $\mathcal{U}([0,1])$

Inverse transform watermark (Kuditipudi et al., 2023)

Let $F(x;\pi) = \sum_{w' \in \mathcal{W}} P_{w'} \cdot \mathbf{1}_{\{\pi(w') \leqslant x\}}$ be the CDF of π-perturbed $\boldsymbol{P}$. Then

$$F^{-1}(U;\pi) = \min\left\{ i : \sum_{w' \in \mathcal{W}} P_{w'} \cdot \mathbf{1}_{\{\pi(w') \leqslant i\}} \geqslant U \right\}$$

with $U \sim \mathcal{U}(0,1)$ satisfies $\pi^{-1}(F^{-1}(U;\pi)) \sim \boldsymbol{P} \equiv (P_w)_{w \in \mathcal{W}}$

$$\mathcal{S}^{\text{inv}}(\boldsymbol{P}, \zeta) := \pi^{-1}(F^{-1}(U;\pi)) \text{ where } \zeta = (U, \pi)$$

- Embedded signal: larger values of U_t tend to correspond to tokens with larger indices

Outline

Human-written vs LLM-generated

Human-written

w_t, ζ_t are *independent*, since a human simply cannot compute ζ_t

LLM-generated

w_t, ζ_t are *dependent* because $w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t)$

Human-written vs LLM-generated

Human-written

w_t, ζ_t are *independent*, since a human simply cannot compute ζ_t

LLM-generated

w_t, ζ_t are *dependent* because $w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t)$

1. Data: $\zeta_t = \mathcal{A}(w_{1:t-1}, \mathrm{Key})$ iid copies of ζ, and tokens $w_1 w_2 \cdots w_n$
2. $\boldsymbol{P}_t$'s are unknown

Human-written vs LLM-generated

Human-written

w_t, ζ_t are *independent*, since a human simply cannot compute ζ_t

LLM-generated

w_t, ζ_t are *dependent* because $w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t)$

1. Data: $\zeta_t = \mathcal{A}(w_{1:t-1}, \mathrm{Key})$ iid copies of ζ, and tokens $w_1 w_2 \cdots w_n$
2. $\boldsymbol{P}_t$'s are unknown

$H_0 : w_{1:n}$ by human

$(w_t, \zeta_t) \mid (w_{1:t-1}, \zeta_{1:t-1}) \stackrel{d}{=} \boldsymbol{P}_t \times \zeta$

$H_1 : w_{1:n}$ by watermarked LLM

$(w_t, \zeta_t) \mid (w_{1:t-1}, \zeta_{1:t-1}) \stackrel{d}{=} (\mathcal{S}(\zeta, \boldsymbol{P}_t), \zeta)$

A challenge: unknown NTP distributions

$$H_0 : w_{1:n} \text{ is by human} \quad vs \quad H_1 : w_{1:n} \text{ is by watermarked LLM}$$

Hypothesis testing

- Under H_0, $(w_t, \zeta_t) \mid (w_{1:t-1}, \zeta_{1:t-1}) \overset{d}{=} \boldsymbol{P}_t \times \zeta$
- Under H_1, $(w_t, \zeta_t) \mid (w_{1:t-1}, \zeta_{1:t-1}) \overset{d}{=} (\mathcal{S}(\zeta, \boldsymbol{P}_t), \zeta)$

A challenge: unknown NTP distributions

$$H_0 : w_{1:n} \text{ is by human} \quad vs \quad H_1 : w_{1:n} \text{ is by watermarked LLM}$$

Hypothesis testing

- Under H_0, $(w_t, \zeta_t) \mid (w_{1:t-1}, \zeta_{1:t-1}) \overset{d}{=} \boldsymbol{P}_t \times \zeta$
- Under H_1, $(w_t, \zeta_t) \mid (w_{1:t-1}, \zeta_{1:t-1}) \overset{d}{=} (\mathcal{S}(\zeta, \boldsymbol{P}_t), \zeta)$

Neyman–Pearson?

A challenge: unknown NTP distributions

$H_0 : w_{1:n}$ is by human $\quad vs \quad H_1 : w_{1:n}$ is by watermarked LLM

Hypothesis testing

- Under H_0, $(w_t, \zeta_t) \mid (w_{1:t-1}, \zeta_{1:t-1}) \stackrel{d}{=} \boldsymbol{P}_t \times \zeta$
- Under H_1, $(w_t, \zeta_t) \mid (w_{1:t-1}, \zeta_{1:t-1}) \stackrel{d}{=} (\mathcal{S}(\zeta, \boldsymbol{P}_t), \zeta)$

Neyman–Pearson? Likelihood ratio:

$$\frac{\mathbb{P}_{H_1}(w_{1:n}, \zeta_{1:n})}{\mathbb{P}_{H_0}(w_{1:n}, \zeta_{1:n})} = \begin{cases} \dfrac{1}{P_{1,w_1} \cdots P_{n,w_n}} & \text{if } \mathcal{S}(\boldsymbol{P}_t, \zeta_t) = w_t \text{ for all } t \\ 0 & \text{otherwise} \end{cases}$$

A challenge: unknown NTP distributions

$$H_0 : w_{1:n} \text{ is by human} \quad vs \quad H_1 : w_{1:n} \text{ is by watermarked LLM}$$

Hypothesis testing

- Under H_0, $(w_t, \zeta_t) \mid (w_{1:t-1}, \zeta_{1:t-1}) \overset{d}{=} \boldsymbol{P}_t \times \zeta$
- Under H_1, $(w_t, \zeta_t) \mid (w_{1:t-1}, \zeta_{1:t-1}) \overset{d}{=} (\mathcal{S}(\zeta, \boldsymbol{P}_t), \zeta)$

Neyman–Pearson? Likelihood ratio:

$$\frac{\mathbb{P}_{H_1}(w_{1:n}, \zeta_{1:n})}{\mathbb{P}_{H_0}(w_{1:n}, \zeta_{1:n})} = \begin{cases} \dfrac{1}{P_{1,w_1} \cdots P_{n,w_n}} & \text{if } \mathcal{S}(\boldsymbol{P}_t, \zeta_t) = w_t \text{ for all } t \\ 0 & \text{otherwise} \end{cases}$$

- But $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_n$ as nuisance are *unknown*, and worse, are varying!

Our approach: pivot under the null

Find a pivotal statistic $Y_t = Y(w_t, \zeta_t)$ such that

- Under H_0, $Y_t \sim \mu_0$, regardless of $\boldsymbol{P}_t$
- Under H_1, $Y_t \sim Y(\mathcal{S}(\zeta_t, \boldsymbol{P}_t), \zeta_t)$, with distribution denoted $\mu_{1,\boldsymbol{P}_t}$

Our approach: pivot under the null

Find a pivotal statistic $Y_t = Y(w_t, \zeta_t)$ such that

- Under H_0, $Y_t \sim \mu_0$, regardless of $\boldsymbol{P}_t$
- Under H_1, $Y_t \sim Y(\mathcal{S}(\zeta_t, \boldsymbol{P}_t), \zeta_t)$, with distribution denoted $\mu_{1,\boldsymbol{P}_t}$

Example: $Y_t = (2w_t - 1)(2\zeta_t - 1) \sim \mathcal{U}(-1, 1)$ for the baby watermark

Our approach: pivot under the null

Find a pivotal statistic $Y_t = Y(w_t, \zeta_t)$ such that

- Under H_0, $Y_t \sim \mu_0$, regardless of $\boldsymbol{P}_t$
- Under H_1, $Y_t \sim Y(\mathcal{S}(\zeta_t, \boldsymbol{P}_t), \zeta_t)$, with distribution denoted $\mu_{1,\boldsymbol{P}_t}$

Example: $Y_t = (2w_t - 1)(2\zeta_t - 1) \sim \mathcal{U}(-1, 1)$ for the baby watermark

Hypothesis testing via pivoting

$$H_0 : Y_t \overset{iid}{\sim} \mu_0,\ t = 1, \ldots, n \qquad \text{vs} \qquad H_1 : Y_t | \boldsymbol{P}_t \sim \mu_{1,\boldsymbol{P}_t},\ t = 1, \ldots, n$$

- Not unique, may lead to information loss, but convenient
- Test distributional difference:

$$T_h = \sum_{t=1}^{n} h(Y_t)$$

for a score function h. Reject H_0 if T_h is larger than a threshold

Pivot for Gumbel-max watermark

Recall $\mathcal{S}^{\mathrm{gum}}(\boldsymbol{P}, \zeta) = \arg\max_{w} \frac{\log U_w}{P_w}$

- A pivotal statistic is $Y_t^{\mathrm{gum}} = U_{t,w_t}$

Pivot for Gumbel-max watermark

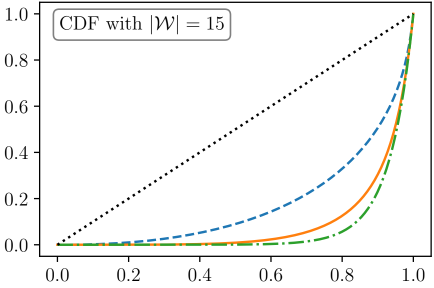
Recall $\mathcal{S}^{\text{gum}}(\boldsymbol{P}, \zeta) = \arg\max_{w} \frac{\log U_w}{P_w}$

- A pivotal statistic is $Y_t^{\text{gum}} = U_{t, w_t}$
  - Under H_0, $Y_t^{\text{gum}} \sim \mathcal{U}(0, 1)$

Pivot for Gumbel-max watermark

Recall $\mathcal{S}^{\mathrm{gum}}(\boldsymbol{P}, \zeta) = \arg\max_{w} \frac{\log U_w}{P_w}$

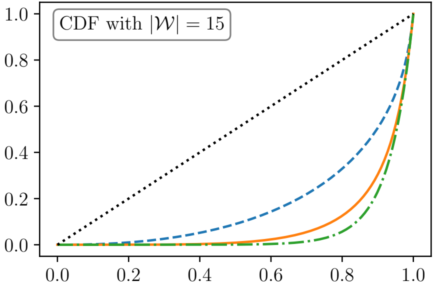
- A pivotal statistic is $Y_t^{\mathrm{gum}} = U_{t,w_t}$
  - Under H_0, $Y_t^{\mathrm{gum}} \sim \mathcal{U}(0,1)$
  - Under H_1, its CDF is $\mathbb{P}_1(Y_t^{\mathrm{gum}} \leqslant r) = \sum_{k=1}^{K} P_{t,k} r^{1/P_{t,k}}$



Pivot for Gumbel-max watermark

Recall $\mathcal{S}^{\text{gum}}(\boldsymbol{P}, \zeta) = \arg\max_{w} \frac{\log U_w}{P_w}$

- A pivotal statistic is $Y_t^{\text{gum}} = U_{t,w_t}$
  - Under H_0, $Y_t^{\text{gum}} \sim \mathcal{U}(0,1)$
  - Under H_1, its CDF is $\mathbb{P}_1(Y_t^{\text{gum}} \leqslant r) = \sum_{k=1}^{K} P_{t,k} r^{1/P_{t,k}}$



Intuition behind this pivot

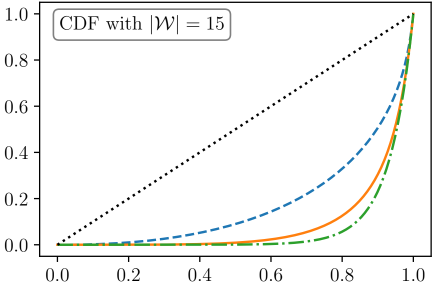
Supremum of likelihood ratio:

$$\sup_{\boldsymbol{P}} \frac{\mathbb{P}_{H_1}(w, \zeta)}{\mathbb{P}_{H_0}(w, \zeta)} = \sup_{\boldsymbol{P}} \frac{\mathbf{1}_{w=\mathcal{S}(\boldsymbol{P}, \zeta)}}{P_w}$$

Pivot for Gumbel-max watermark

Recall $\mathcal{S}^{\text{gum}}(\boldsymbol{P}, \zeta) = \arg\max_w \frac{\log U_w}{P_w}$

- A pivotal statistic is $Y_t^{\text{gum}} = U_{t,w_t}$
  - Under H_0, $Y_t^{\text{gum}} \sim \mathcal{U}(0,1)$
  - Under H_1, its CDF is $\mathbb{P}_1(Y_t^{\text{gum}} \leqslant r) = \sum_{k=1}^{K} P_{t,k} r^{1/P_{t,k}}$



Intuition behind this pivot

Supremum of likelihood ratio:

$$\sup_{\boldsymbol{P}} \frac{\mathbb{P}_{H_1}(w,\zeta)}{\mathbb{P}_{H_0}(w,\zeta)} = \sup_{\boldsymbol{P}} \frac{\mathbf{1}_{w=\mathcal{S}(\boldsymbol{P},\zeta)}}{P_w}$$

- Asymptotically determined by U_w

Pivot for inverse transform watermark

- Recall that $\zeta_t = (\pi_t, U_t) \sim$ uniform permutations $\times \, \mathcal{U}(0,1)$. Define $\eta(k) = (k-1)/(K-1)$
- A pivotal statistic is $Y_t^{\text{dif}} = |U_t - \eta(\pi_t(w_t))|$ (Kuditipudi et al., 2023)
- Under H_0,

$$\lim_{|\mathcal{W}| \to \infty} \mathbb{P}_{H_0}(Y_t^{\text{dif}} \leqslant r) = 1 - (1-r)^2 \ \text{ for any } \ r \in [0,1]$$

Outline

What's the right notion of statistical efficiency?

Class-dependent statistical efficiency

Fixing Type I error, a watermark is preferred if it has a higher power

- Comparison depends on $\boldsymbol{P}_t$'s

Class-dependent statistical efficiency

Fixing Type I error, a watermark is preferred if it has a higher power

- Comparison depends on $\boldsymbol{P}_t$'s
- Minimax viewpoint: unfortunately, all watermarks are powerless over all NTP distributions

Class-dependent statistical efficiency

Fixing Type I error, a watermark is preferred if it has a higher power

- Comparison depends on $\boldsymbol{P}_t$'s
- Minimax viewpoint: unfortunately, all watermarks are powerless over all NTP distributions

Class-dependent efficiency

- Find structured $\mathcal{P}$ that contains all NTP distributions $\boldsymbol{P}_t$
- Find the lowest power over $\mathcal{P}$

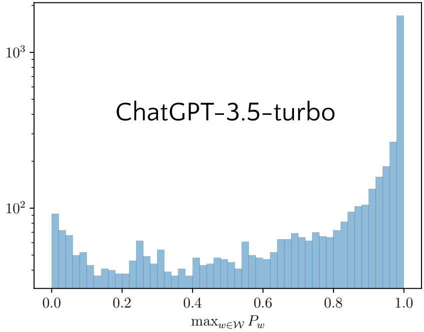
Class-dependent statistical efficiency

Fixing Type I error, a watermark is preferred if it has a higher power

- Comparison depends on $\boldsymbol{P}_t$'s
- Minimax viewpoint: unfortunately, all watermarks are powerless over all NTP distributions

Class-dependent efficiency

- Find structured $\mathcal{P}$ that contains all NTP distributions $\boldsymbol{P}_t$
- Find the lowest power over $\mathcal{P}$



A class of NTP distributions

Δ-regular distribution class

$$\mathcal{P}_\Delta := \{\boldsymbol{P} = (P_1, \cdots, P_k) : \max_k P_k \leqslant 1 - \Delta\}$$

- Chopping off *deterministic* NTP distributions of the form $(0, \ldots, 0, 1, 0, \ldots, 0)$
- Shannon entropy satisfies

$$\mathrm{Ent}(\boldsymbol{P}) = \sum P_w \log \frac{1}{P_w} \geqslant \sum P_w(1 - P_w) \geqslant \sum P_w \cdot \Delta = \Delta$$

A detour: why you can start doing watermark research even today

You don’t need GPUs to work on watermarks!

```
import tiktoken
import openai
import math
import numpy as np
from tqdm import tqdm
import os
from IPython import embed
import nltk
from nltk import tokenize
nltk.download('punkt')
from statsmodels.distributions.empirical_distribution import ECDF
import matplotlib
matplotlib.use('Agg')
import matplotlib.pyplot as plt
plt.rcParams.update({
    'font.size': 12,
    'text.usetex': True,
    'text.latex.preamble': r'\usepackage{amsfonts}'
})
```

```
[nltk_data] Downloading package punkt to /Users/lixiang/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

You don't need GPUs to work on watermarks!

```
[ ]  ## Token info

     openai.api_key = 'Please input your OpenAI key here'
     # print(openai.Model.list())

     # model = "text-davinci-003"
     # model = "gpt-4"
     model = "gpt-3.5-turbo-instruct"
     tokens = ["Yes", "No"]
     tokenizer = tiktoken.encoding_for_model(model)
     ids = [tokenizer.encode(token) for token in tokens]
     yes_id = ids[0][0]
     no_id = ids[1][0]
```

```
[ ]  def get_completion(prompt, temp=0.):
         response = openai.Completion.create(model=model,
                                             prompt=prompt,
                                             max_tokens=1000,
                                             temperature=temp,
                                             logprobs=5)
         return response
```

```
[ ]  a = get_completion("what you name", temp=0.)
```

Asymptotic class-dependent efficiency

Theorem

Fixing Type I error in $(0,1)$*, the pivot-based test statistic* $T_h = \sum h(Y_t)$ *satisfies*

$$\limsup_{n\to\infty} \textit{Type II error}^{\frac{1}{n}} \leqslant \exp(-R_{\mathcal{P}}(h)),$$

where $\mathcal{P}$*-efficiency rate* $R_{\mathcal{P}}(h)$ *is*

$$R_{\mathcal{P}}(h) = -\inf_{\theta\geqslant 0}\{\theta \mathbb{E}_0 h(Y) + \log\phi_{\mathcal{P},h}(\theta)\} \quad \textit{with} \quad \phi_{\mathcal{P},h}(\theta) = \sup_{\boldsymbol{P}\in\mathcal{P}} \mathbb{E}_{1,\boldsymbol{P}}\, \mathrm{e}^{-\theta h(Y)}$$

Asymptotic class-dependent efficiency

Theorem

Fixing Type I error in $(0,1)$*, the pivot-based test statistic* $T_h = \sum h(Y_t)$ *satisfies*

$$\limsup_{n\to\infty} \textit{Type II error}^{\frac{1}{n}} \leqslant \exp(-R_{\mathcal{P}}(h)),$$

where $\mathcal{P}$*-efficiency rate* $R_{\mathcal{P}}(h)$ *is*

$$R_{\mathcal{P}}(h) = -\inf_{\theta\geqslant 0}\{\theta\mathbb{E}_0 h(Y) + \log\phi_{\mathcal{P},h}(\theta)\} \quad \textit{with} \quad \phi_{\mathcal{P},h}(\theta) = \sup_{\boldsymbol{P}\in\mathcal{P}} \mathbb{E}_{1,\boldsymbol{P}}\, \mathrm{e}^{-\theta h(Y)}$$

- Tight in the minimax sense. Bahadur efficiency when $\mathcal{P}$ is a singleton

Asymptotic class-dependent efficiency

Theorem

Fixing Type I error in $(0,1)$*, the pivot-based test statistic* $T_h = \sum h(Y_t)$ *satisfies*

$$\limsup_{n\to\infty} \textit{Type II error}^{\frac{1}{n}} \leqslant \exp(-R_{\mathcal{P}}(h)),$$

where $\mathcal{P}$*-efficiency rate* $R_{\mathcal{P}}(h)$ *is*

$$R_{\mathcal{P}}(h) = -\inf_{\theta\geqslant 0}\left\{\theta\mathbb{E}_0 h(Y) + \log\phi_{\mathcal{P},h}(\theta)\right\} \quad \textit{with} \quad \phi_{\mathcal{P},h}(\theta) = \sup_{\boldsymbol{P}\in\mathcal{P}} \mathbb{E}_{1,\boldsymbol{P}}\, \mathrm{e}^{-\theta h(Y)}$$

- Tight in the minimax sense. Bahadur efficiency when $\mathcal{P}$ is a singleton
- Monotonicity: $R_{\mathcal{P}_1}(h) \geqslant R_{\mathcal{P}_2}(h)$ if $\mathcal{P}_1 \subset \mathcal{P}_2$

Asymptotic class-dependent efficiency

Theorem

Fixing Type I error in $(0,1)$*, the pivot-based test statistic* $T_h = \sum h(Y_t)$ *satisfies*

$$\limsup_{n\to\infty} \text{Type II error}^{\frac{1}{n}} \leqslant \exp(-R_{\mathcal{P}}(h)),$$

where $\mathcal{P}$*-efficiency rate* $R_{\mathcal{P}}(h)$ *is*

$$R_{\mathcal{P}}(h) = -\inf_{\theta \geqslant 0} \{\theta \mathbb{E}_0 h(Y) + \log \phi_{\mathcal{P},h}(\theta)\} \quad \textit{with} \quad \phi_{\mathcal{P},h}(\theta) = \sup_{\boldsymbol{P} \in \mathcal{P}} \mathbb{E}_{1,\boldsymbol{P}}\, \mathrm{e}^{-\theta h(Y)}$$

- Tight in the minimax sense. Bahadur efficiency when $\mathcal{P}$ is a singleton
- Monotonicity: $R_{\mathcal{P}_1}(h) \geqslant R_{\mathcal{P}_2}(h)$ if $\mathcal{P}_1 \subset \mathcal{P}_2$
- $R_{\mathcal{P}}(h) = 0$ for any h if $\mathcal{P}$ includes $(0, \ldots, 0, 1, 0, \ldots, 0)$, thereby justifying $\mathcal{P}_\Delta$

Efficiency of the baby watermark

Theorem

$$\limsup_{n\to\infty} \textit{Type II error}^{\frac{1}{n}} \leqslant \exp(-R_{\mathcal{P}}(h)),$$

where

$$R_{\mathcal{P}}(h) = -\inf_{\theta\geqslant 0}\{\theta\mathbb{E}_0 h(Y) + \log\phi_{\mathcal{P},h}(\theta)\} \ \ \textit{with} \ \ \phi_{\mathcal{P},h}(\theta) = \sup_{\boldsymbol{P}\in\mathcal{P}} \mathbb{E}_{1,\boldsymbol{P}}\, \mathrm{e}^{-\theta h(Y)}$$

Let $\mathcal{W} = \{0,1\}$, $\boldsymbol{P}_t = (P_{t,0}, P_{t,1})$, ζ_t be iid copies of $\mathcal{U}(0,1)$, with decoder

$$w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t) = \begin{cases} 0 & \text{if } \zeta_t \leqslant P_{t,0} \\ 1 & \text{otherwise} \end{cases}$$

and pivot $Y(w_t, \zeta_t) = (2w_t - 1)(2\zeta_t - 1)$. With h being identity, $R_{\mathcal{P}_\Delta}(h)$ is

$$-\inf_{\theta\geqslant 0} \log\frac{1}{\theta}\left[\frac{\mathrm{e}^{\theta(1-2\Delta)} + \mathrm{e}^{-\theta(1-2\Delta)}}{2} - \mathrm{e}^{-\theta}\right]$$

A minimax formulation for $R_{\mathcal{P}}$

$$R_{\mathcal{P}}(h) = -\inf_{\theta \geqslant 0} \left\{ \theta \mathbb{E}_0 h(Y) + \sup_{\boldsymbol{P} \in \mathcal{P}} \log \left(\mathbb{E}_{1,\boldsymbol{P}} \, \mathrm{e}^{-\theta h(Y)} \right) \right\}$$

Finding the optimal score $h^\star = \arg\max_h R_{\mathcal{P}}(h)$ reduces to a minimax problem:

$$\min_h \max_{\boldsymbol{P} \in \mathcal{P}} L(h, \boldsymbol{P}) \quad \text{where} \quad L(h, \boldsymbol{P}) := \mathbb{E}_0 h(Y) + \log \left(\mathbb{E}_{1,\boldsymbol{P}} \, \mathrm{e}^{-h(Y)} \right)$$

A minimax formulation for $R_{\mathcal{P}}$

$$R_{\mathcal{P}}(h) = -\inf_{\theta \geqslant 0} \left\{ \theta \mathbb{E}_0 h(Y) + \sup_{\boldsymbol{P} \in \mathcal{P}} \log \left(\mathbb{E}_{1,\boldsymbol{P}} \, \mathrm{e}^{-\theta h(Y)} \right) \right\}$$

Finding the optimal score $h^\star = \arg\max_h R_{\mathcal{P}}(h)$ reduces to a minimax problem:

$$\min_h \max_{\boldsymbol{P} \in \mathcal{P}} L(h, \boldsymbol{P}) \quad \text{where} \quad L(h, \boldsymbol{P}) := \mathbb{E}_0 h(Y) + \log \left(\mathbb{E}_{1,\boldsymbol{P}} \, \mathrm{e}^{-h(Y)} \right)$$

- Unfortunately, the minimax problem is generally not convex-concave

A minimax formulation for $R_{\mathcal{P}}$

$$R_{\mathcal{P}}(h) = -\inf_{\theta \geqslant 0} \left\{ \theta \mathbb{E}_0 h(Y) + \sup_{\boldsymbol{P} \in \mathcal{P}} \log \left(\mathbb{E}_{1,\boldsymbol{P}} \, \mathrm{e}^{-\theta h(Y)} \right) \right\}$$

Finding the optimal score $h^\star = \arg\max_h R_{\mathcal{P}}(h)$ reduces to a minimax problem:

$$\min_{h} \max_{\boldsymbol{P} \in \mathcal{P}} L(h, \boldsymbol{P}) \text{ where } L(h, \boldsymbol{P}) := \mathbb{E}_0 h(Y) + \log \left(\mathbb{E}_{1,\boldsymbol{P}} \, \mathrm{e}^{-h(Y)} \right)$$

- Unfortunately, the minimax problem is generally not convex-concave
- Case-by-case analysis is required, but we are often lucky

Outline

1. Preliminaries

2. Hypothesis testing formulation

3. Efficiency and optimal detection

4. Application to Gumbel-max watermark

5. Robust detection

Analysis of the Gumbel-max watermark

$$\mathcal{S}^{\text{gum}}(\boldsymbol{P}, \zeta) = \arg\max_{w} \frac{\log U_w}{P_w} \text{ where } \zeta = (U_1, \ldots, U_K)$$

with pivot $Y^{\text{gum}} = U_{t,w_t}$

Lemma (Convexity lemma)

For any non-decreasing function h, the following is a convex function in $\boldsymbol{P}$:

$$\boldsymbol{P} \mapsto \phi_h(\boldsymbol{P}) := \mathbb{E}_{1,\boldsymbol{P}}\, \mathrm{e}^{-h(Y^{\text{gum}})}$$

- Max part of $\min_{h} \max_{\boldsymbol{P} \in \mathcal{P}} L(h, \boldsymbol{P})$ is

$$\sup_{\boldsymbol{P} \in \mathcal{P}} \log \left(\mathbb{E}_{1,\boldsymbol{P}},\, \mathrm{e}^{-h(Y^{\text{gum}})} \right) = \log \sup_{\boldsymbol{P} \in \mathcal{P}} \phi_h(\boldsymbol{P})$$

Analysis of the Gumbel-max watermark

$$\mathcal{S}^{\text{gum}}(\boldsymbol{P}, \zeta) = \arg\max_{w} \frac{\log U_w}{P_w} \text{ where } \zeta = (U_1, \ldots, U_K)$$

with pivot $Y^{\text{gum}} = U_{t,w_t}$

Lemma (Convexity lemma)

For any non-decreasing function h, the following is a convex function in $\boldsymbol{P}$:

$$\boldsymbol{P} \mapsto \phi_h(\boldsymbol{P}) := \mathbb{E}_{1,\boldsymbol{P}} \, \mathrm{e}^{-h(Y^{\text{gum}})}$$

- Max part of $\min_h \max_{\boldsymbol{P} \in \mathcal{P}} L(h, \boldsymbol{P})$ is

$$\sup_{\boldsymbol{P} \in \mathcal{P}} \log \left(\mathbb{E}_{1,\boldsymbol{P}}, \mathrm{e}^{-h(Y^{\text{gum}})} \right) = \log \sup_{\boldsymbol{P} \in \mathcal{P}} \phi_h(\boldsymbol{P})$$

- Maximizing a convex function over a *convex* set requires examining only the extreme points!

Analysis of the Gumbel-max watermark

Lemma (Convexity lemma)

For any non-decreasing function h, the following is a convex function in $\boldsymbol{P}$:

$$\boldsymbol{P} \mapsto \phi_h(\boldsymbol{P}) := \mathbb{E}_{1,\boldsymbol{P}} \, \mathrm{e}^{-h(Y^{\mathrm{gum}})}$$

- Extreme points of $\mathcal{P}_\Delta$ are

$$\boldsymbol{P}_\Delta^\star = \Big(\underbrace{1-\Delta, \ldots, 1-\Delta}_{\lfloor \frac{1}{1-\Delta} \rfloor \text{ times}}, \widetilde{\Delta}, 0, \ldots\Big) \text{ with } \widetilde{\Delta} = 1 - (1-\Delta) \cdot \left\lfloor \frac{1}{1-\Delta} \right\rfloor$$

  and all its permutations

Proof sketch of the convexity lemma I

Lemma (Convexity lemma)

For any non-decreasing function h, the following is a convex function in $\boldsymbol{P}$:

$$\boldsymbol{P} \mapsto \phi_h(\boldsymbol{P}) := \mathbb{E}_{1,\boldsymbol{P}}, \mathrm{e}^{-h(Y^{\mathrm{gum}})}$$

- Y^{gum} has a mixture of Beta distributions:

$$F_{1,\boldsymbol{P}}(r) = \sum_{w \in \mathcal{W}} P_w r^{1/P_w}$$

Proof sketch of the convexity lemma II

- Show that $\boldsymbol{P} \mapsto F_{1,\boldsymbol{P}}(r)$ is convexfor any given $r \in [0,1]$:

$$\nabla^2_{\boldsymbol{P}} F_{1,\boldsymbol{P}}(r) = \begin{bmatrix} r^{1/P_1} \frac{\log^2 r}{P_1^3} & 0 & \dots & 0 \\ 0 & r^{1/P_2} \frac{\log^2 r}{P_2^3} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & r^{1/P_{|\mathcal{W}|}} \frac{\log^2 r}{P_{|\mathcal{W}|}^3} \end{bmatrix} \succeq 0$$

- $\phi_h(\boldsymbol{P})$ is a nonnegative weighted sum of $F_{1,\boldsymbol{P}}(r)$:

$$\begin{aligned} \phi_h(\boldsymbol{P}) = F_{1,\boldsymbol{P}}(r) \mathrm{e}^{-h(r)} \Big|_0^1 + \int_0^1 F_{1,\boldsymbol{P}}(r) \mathrm{e}^{-h(r)} h(\mathrm{d}r) \\ = \mathrm{e}^{-h(1)} + \int_0^1 F_{1,\boldsymbol{P}}(r) \mathrm{e}^{-h(r)} h(\mathrm{d}r) \end{aligned}$$

Find optimal detection for Gumbel-max watermark

For non-decreasing h, we have $\sup_{\boldsymbol{P} \in \mathcal{P}_\Delta} \mathbb{E}_{1,\boldsymbol{P}}\, \mathrm{e}^{-h(Y^{\mathrm{gum}})} = \mathbb{E}_{1,\boldsymbol{P}_\Delta^\star}\, \mathrm{e}^{-h(Y^{\mathrm{gum}})}$

Denoting by $\boldsymbol{P}_\Delta^\star$ any vertex (extreme point) of $\mathcal{P}_\Delta$. For any h,

$$
\begin{aligned}
&\min_h \max_{\boldsymbol{P} \in \mathcal{P}_\Delta} \mathbb{E}_0 h(Y^{\mathrm{gum}}) + \log\left(\mathbb{E}_{1,\boldsymbol{P}}\, \mathrm{e}^{-h(Y^{\mathrm{gum}})}\right) \\
&\geqslant \min_h \mathbb{E}_0 h(Y^{\mathrm{gum}}) + \log\left(\mathbb{E}_{1,\boldsymbol{P}_\Delta^\star}\, \mathrm{e}^{-h(Y^{\mathrm{gum}})}\right) \\
&= -D_{\mathrm{KL}}(\mu_0, \mu_{1,\boldsymbol{P}_\Delta^\star}),
\end{aligned}
$$

where the equality follows from the Donsker–Varadhan representation, attained at $h = h^\star := \log \frac{\mathrm{d}\mu_{1,\boldsymbol{P}_\Delta^\star}}{\mathrm{d}\mu_0}$

When $h = h^\star$ the inequality reduces to equality, because it is non-decreasing

Optimal detection for Gumbel-max watermark

Theorem

The optimal score function that achieves the highest $\mathcal{P}_\Delta$-efficiency rate $R_{\mathcal{P}_\Delta}(h)$ takes the form

$$h^\star_{\mathrm{gum},\Delta}(y) = \log\left(\left\lfloor \frac{1}{1-\Delta} \right\rfloor y^{\frac{\Delta}{1-\Delta}} + y^{\frac{\tilde{\Delta}}{1-\tilde{\Delta}}}\right), \quad \textit{with} \quad \widetilde{\Delta} = (1-\Delta)\left\lfloor \frac{1}{1-\Delta} \right\rfloor$$

Optimal detection for Gumbel-max watermark

Theorem

The optimal score function that achieves the highest $\mathcal{P}_\Delta$-efficiency rate $R_{\mathcal{P}_\Delta}(h)$ takes the form

$$h^\star_{\mathrm{gum},\Delta}(y) = \log\left(\left\lfloor \frac{1}{1-\Delta} \right\rfloor y^{\frac{\Delta}{1-\Delta}} + y^{\frac{\widetilde{\Delta}}{1-\widetilde{\Delta}}}\right), \ \textit{with} \ \widetilde{\Delta} = (1-\Delta)\left\lfloor \frac{1}{1-\Delta} \right\rfloor$$

- $h^\star_{\mathrm{gum},\Delta} = h^\star = \log \dfrac{\mathrm{d}\mu_{1,\boldsymbol{P}^\star_\Delta}}{\mathrm{d}\mu_0}$

Optimal detection for Gumbel-max watermark

Theorem

The optimal score function that achieves the highest $\mathcal{P}_\Delta$-efficiency rate $R_{\mathcal{P}_\Delta}(h)$ takes the form

$$h^\star_{\mathrm{gum},\Delta}(y) = \log\left(\left\lfloor \frac{1}{1-\Delta} \right\rfloor y^{\frac{\Delta}{1-\Delta}} + y^{\frac{\tilde{\Delta}}{1-\tilde{\Delta}}} \right), \text{ with } \widetilde{\Delta} = (1-\Delta)\left\lfloor \frac{1}{1-\Delta} \right\rfloor$$

- $h^\star_{\mathrm{gum},\Delta} = h^\star = \log \frac{\mathrm{d}\mu_{1,\boldsymbol{P}^\star_\Delta}}{\mathrm{d}\mu_0}$
- Aaronson (2023) proposed $h_{\mathrm{ars}}(y) = -\log(1-y)$
- Kuditipudi et al. (2023); Fernandez et al. (2023) proposed $h_{\log}(y) = \log y$

Comparison with other detection rules

Theorem

There exists an absolute constant $\Delta^\star \approx 0.17756$ such that the following two statements hold:

(a) *When $0.001 < \Delta < \Delta^\star$, h_{ars} has higher $\mathcal{P}_\Delta$-efficiency than $h_{\log}$:*

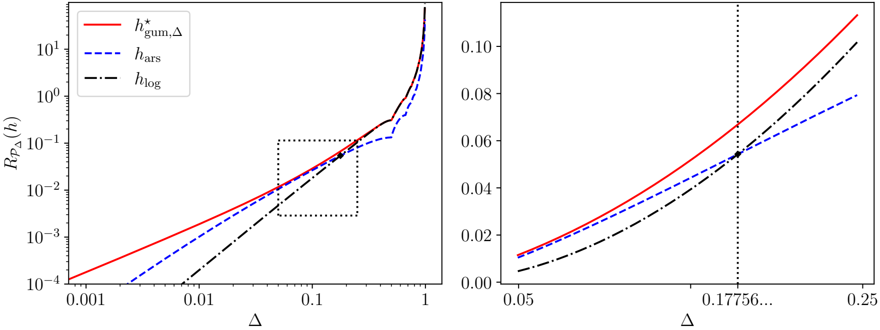
$$R_{\mathcal{P}_\Delta}(h_{\log}) < R_{\mathcal{P}_\Delta}(h_{\text{ars}}) < R_{\mathcal{P}_\Delta}(h^\star_{\text{gum},\Delta})$$

(b) *When $\Delta^\star < \Delta < 0.99$, $h_{\log}$ has higher $\mathcal{P}_\Delta$-efficiency than h_{ars}:*

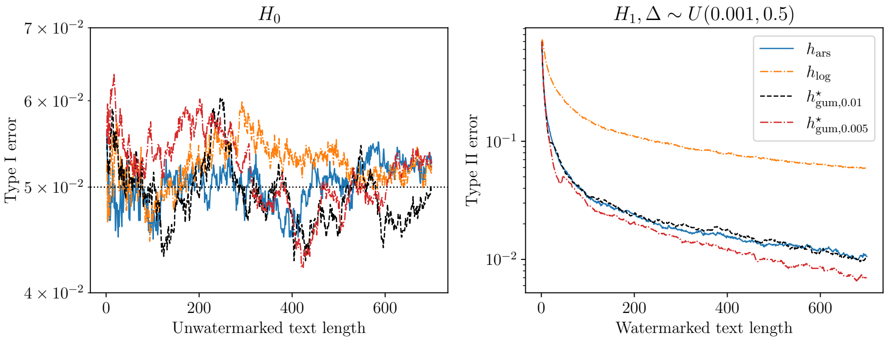
$$R_{\mathcal{P}_\Delta}(h_{\text{ars}}) < R_{\mathcal{P}_\Delta}(h_{\log}) < R_{\mathcal{P}_\Delta}(h^\star_{\text{gum},\Delta})$$

- In any case, $h^\star_{\text{gum},\Delta}$ has the highest rate

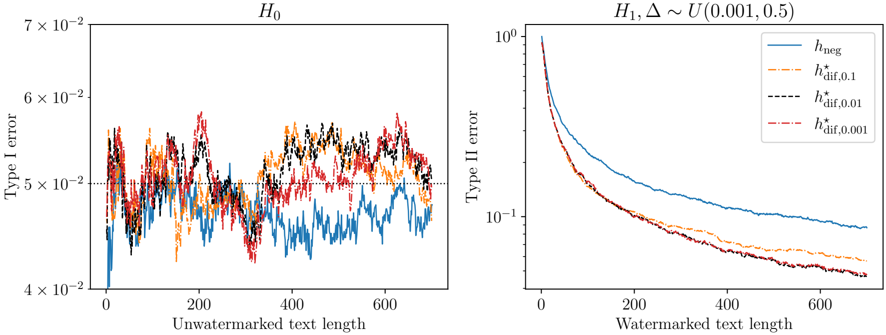
Illustration of the superiority of $h^\star_{\text{gum},\Delta}$



Numerical results for Gumbel-max watermark

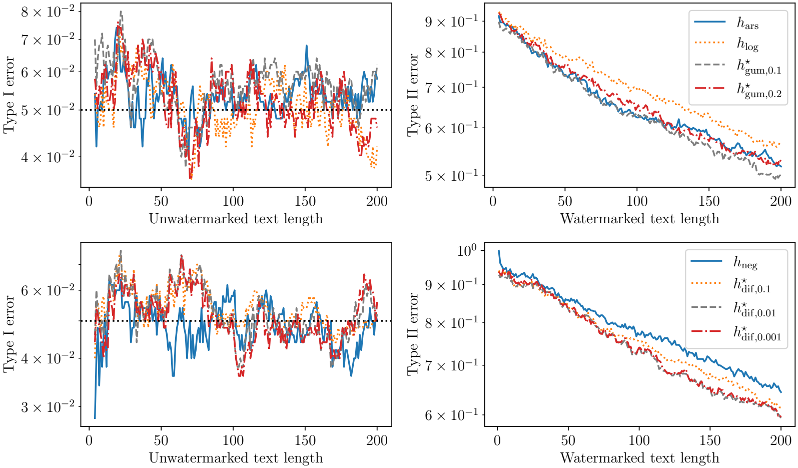


Numerical results for inverse transform watermark



Experiments on the C4 dataset using OPT-1.3B

Left: Type I; Right: Type II; Top: Gumbel-max; Bottom: Inverse transform



Outline

1. Preliminaries

2. Hypothesis testing formulation

3. Efficiency and optimal detection

4. Application to Gumbel-max watermark

5. Robust detection

Watermark under text modification

A student might modify the text generated from an LLM, either due to personalization or to try to escape from detection

Watermark under text modification

A student might modify the text generated from an LLM, either due to personalization or to try to escape from detection

- To cope with modification, Gumbel-max watermark uses a few tokens to compute pseudorandom numbers

  For example, $\zeta_t = \mathcal{A}(w_{t-5:t-1}, \mathrm{Key})$, using the last 5 tokens

- A modified token will turn the watermark signals in the next few 5 tokens to noise

Watermark under text modification

A student might modify the text generated from an LLM, either due to personalization or to try to escape from detection

- To cope with modification, Gumbel-max watermark uses a few tokens to compute pseudorandom numbers

  For example, $\zeta_t = \mathcal{A}(w_{t-5:t-1}, \mathrm{Key})$, using the last 5 tokens

- A modified token will turn the watermark signals in the next few 5 tokens to noise

Hypothesis testing under mixtures

$$H_0 : Y_t \sim \mu_0 \quad \text{vs} \quad H_1^{\mathrm{mix}} : Y_t | \boldsymbol{P}_t \sim (1-\eta_t)\mu_0 + \eta_t \mu_{1,\boldsymbol{P}_t}$$

Watermark under text modification

A student might modify the text generated from an LLM, either due to personalization or to try to escape from detection

- To cope with modification, Gumbel-max watermark uses a few tokens to compute pseudorandom numbers

  For example, $\zeta_t = \mathcal{A}(w_{t-5:t-1}, \mathrm{Key})$, using the last 5 tokens

- A modified token will turn the watermark signals in the next few 5 tokens to noise

Hypothesis testing under mixtures

$$H_0 : Y_t \sim \mu_0 \quad \text{vs} \quad H_1^{\mathrm{mix}} : Y_t | \boldsymbol{P}_t \sim (1-\eta_t)\mu_0 + \eta_t \mu_{1,\boldsymbol{P}_t}$$

- $\eta_t \in \{0, 1\}$ is independent or modeled by a Markov process
- Sparse mixture detection

When is detection statistically possible?

The large deviation regime ($\eta_t = 1$ and $\Delta > 0$ constant) is too easy

A (difficulty) scaling regime

- $\mathbb{E}\eta_t = \varepsilon_n$ with $\varepsilon_n \asymp n^{-p}$ for $p \in (0, 1]$
- $\max_{w \in \mathcal{W}} \boldsymbol{P}_{t,w} = 1 - \Delta_n$ with $\Delta_n \asymp n^{-q}$ for $q \in (0, 1)$

Theorem (Phase transition)

- *If* $q + 2p > 1$, H_0 *and* H_1^{mix} *merge asymptotically*
- *If* $q + 2p < 1$, H_0 *and* H_1^{mix} *separate asymptotically*

- How to achieve robust detection in the regime $q + 2p < 1$? LRT is impractical since it requires knowing $\boldsymbol{P}_t$'s

Optimal adaptive detection: Goodness-of-fit (GoF) test

- Empirical CDF of p-values: $\mathbb{F}_n(r) = \frac{1}{n}\sum_{t=1}^{n} 1_{\mathsf{p}_t \leqslant r}$ where $\mathsf{p}_t = 1 - Y_t^{\mathrm{gum}}$
- Introduce a scalar convex function indexed by s:

$$\phi_s(x) = \begin{cases} x\log x - x + 1, & \text{if } s = 1 \\ \dfrac{1 - s + sx - x^s}{s(1-s)}, & \text{if } s \neq 0, 1 \\ -\log x + x - 1, & \text{if } s = 0 \end{cases}$$

- ϕ_s-divergence between $\mathrm{Bern}(u)$ and $\mathrm{Bern}(v)$:

$$K_s(u, v) = v\phi_s\left(\frac{u}{v}\right) + (1 - v)\phi_s\left(\frac{1-u}{1-v}\right)$$

- For $s \in [0, 2]$, reject H_0 if $nS_n^+(s) := n \sup_{r\in(0,1)} K_s(\mathbb{F}_n(r), r) 1_{\mathbb{F}_n(r) > r}$ is larger than a certain threshold

Adaptive optimality and optimal efficiency

Theorem (Adaptive optimality)

Let $q + 2p < 1$ *and* $s \in [0, 2]$*. Setting the threshold* $\asymp \log \log n$*, both the Type I and II errors of the GoF test tend to 0 as* $n \to \infty$

Adaptive optimality and optimal efficiency

Theorem (Adaptive optimality)

Let $q + 2p < 1$ and $s \in [0, 2]$. Setting the threshold $\asymp \log\log n$, both the Type I and II errors of the GoF test tend to 0 as $n \to \infty$

- Optimality without any prior knowledge

Adaptive optimality and optimal efficiency

Theorem (Adaptive optimality)

Let $q + 2p < 1$ and $s \in [0, 2]$. Setting the threshold $\asymp \log\log n$, both the Type I and II errors of the GoF test tend to 0 as $n \to \infty$

- Optimality without any prior knowledge

Optimal efficiency

Let $s \in (0, 1)$, $\varepsilon_n \equiv \varepsilon \in (0, 1]$ and $\Delta_n \equiv \Delta \in (0, 1)$. The score function $S_n^+(s)$ has

$$R_{\mathcal{P}_\Delta}(S_n^+(s)) = \sup_{\text{measurable } S_n} R_{\mathcal{P}_\Delta}(S_n) = D_{\mathrm{KL}}(\mu_0, (1-\varepsilon)\mu_0 + \varepsilon\mu_{1,\boldsymbol{P}_\Delta^\star})$$

- When $\varepsilon = 1$, this rate is obtained by the sum-based test based on $h^\star_{\mathrm{gum},\Delta}$

Adaptive optimality and optimal efficiency

Theorem (Adaptive optimality)

Let $q + 2p < 1$ and $s \in [0, 2]$. Setting the threshold $\asymp \log\log n$, both the Type I and II errors of the GoF test tend to 0 as $n \to \infty$

- Optimality without any prior knowledge

Optimal efficiency

Let $s \in (0, 1)$, $\varepsilon_n \equiv \varepsilon \in (0, 1]$ and $\Delta_n \equiv \Delta \in (0, 1)$. The score function $S_n^+(s)$ has

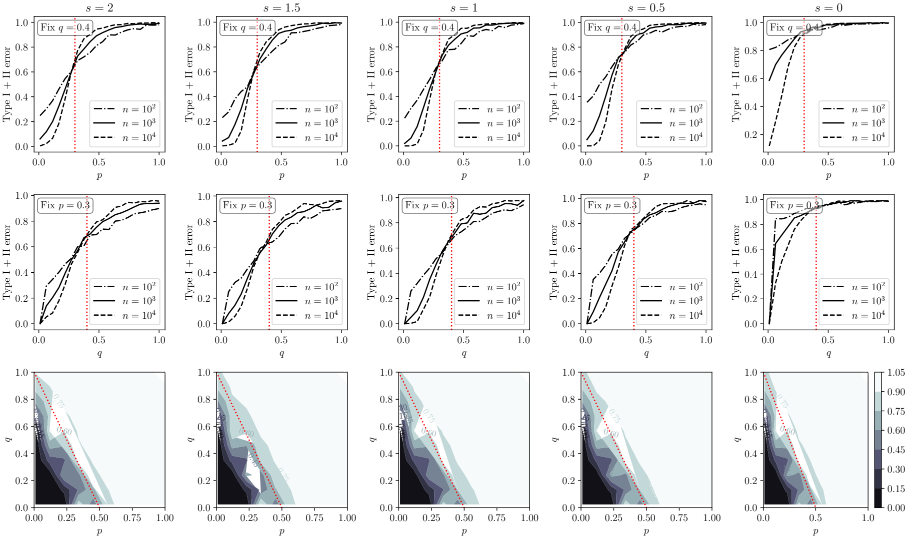
$$R_{\mathcal{P}_\Delta}(S_n^+(s)) = \sup_{\text{measurable } S_n} R_{\mathcal{P}_\Delta}(S_n) = D_{\mathrm{KL}}(\mu_0, (1 - \varepsilon)\mu_0 + \varepsilon\mu_{1,\boldsymbol{P}_\Delta^\star})$$

- When $\varepsilon = 1$, this rate is obtained by the sum-based test based on $h^\star_{\mathrm{gum},\Delta}$

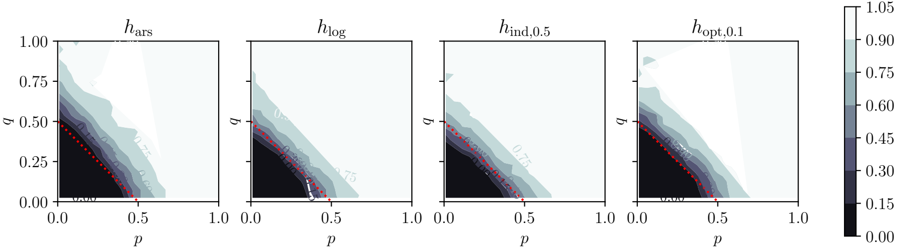
Theorem (Suboptimality of sum-based tests)

When $\varepsilon < 1$, the detection boundary for sum-based tests is $p + q = 1/2$ for the Gumbel-max watermark

Empirical detection boundaries



Suboptimality of sum-based tests



Concluding remarks

Take-home messages

- *A Statistical Framework of Watermarks for Large Language Models: Pivot, Detection Efficiency and Optimal Rules*. The Annals of Statistics, 2025
- *Robust Detection of Watermarks for Large Language Models Under Human Edits*. arXiv:2411.13868

- A statistical framework for (unbiased) watermarks of LLMs
- Defined class-dependent efficiency measure to evaluate detection
- Identified the optimal detection rule according to the efficiency measure
- Achieved adaptive optimality for robust estimation using GoF tests

Future directions

- Extend the analysis to finite-sample
- Multiple testing in the case of multiple LLMs (ChatGPT, Claude, ...)?
- Investigate data-driven distribution classes
- ......

References I

S. Aaronson. Watermarking of large language models, August 2023. URL https://simons.berkeley.edu/talks/scott-aaronson-ut-austin-openai-2023-08-17.

G. Bao, Y. Zhao, Z. Teng, L. Yang, and Y. Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*, 2023.

M. Christ, S. Gunn, and O. Zamir. Undetectable watermarks for language models. *arXiv preprint arXiv:2306.09194*, 2023.

P. Fernandez, A. Chaffin, K. Tit, V. Chappelier, and T. Furon. Three bricks to consolidate watermarks for large language models. *arXiv preprint arXiv:2308.00113*, 2023.

GPTZero. GPTZero: More than an AI detector preserve what's human. https://gptzero.me/, 2023.

Z. Hu, L. Chen, X. Wu, Y. Wu, H. Zhang, and H. Huang. Unbiased watermark for large language models. *arXiv preprint arXiv:2310.10669*, 2023.

J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, volume 202, pages 17061–17084, 2023a.

J. Kirchenbauer, J. Geiping, Y. Wen, M. Shu, K. Saifullah, K. Kong, K. Fernando, A. Saha, M. Goldblum, and T. Goldstein. On the reliability of watermarks for large language models. *arXiv preprint arXiv:2306.04634*, 2023b.

K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

R. Kuditipudi, J. Thickstun, T. Hashimoto, and P. Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.

W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou. GPT detectors are biased against non-native english writers. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023.

W. Liang, Z. Izzo, Y. Zhang, H. Lepp, H. Cao, X. Zhao, L. Chen, H. Ye, S. Liu, Z. Huang, et al. Monitoring AI-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*, 2024.

References II

E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*, 2023.

V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

D. Weber-Wulff, A. Anohina-Naumeca, S. Bjelobaba, T. Foltỳnek, J. Guerrero-Dib, O. Popoola, P. **v**Sigut, and L. Waddington. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1):26, 2023.

Y. Wu, Z. Hu, H. Zhang, and H. Huang. DiPmark: A stealthy, efficient and resilient watermark for large language models. *arXiv preprint arXiv:2310.07710*, 2023.

X. Yang, W. Cheng, L. Petzold, W. Y. Wang, and H. Chen. DNA-GPT: Divergent n-gram analysis for training-free detection of GPT-generated text. *arXiv preprint arXiv:2305.17359*, 2023.

ZeroGPT. ZeroGPT: Trusted GPT-4, ChatGPT and AI detector tool by ZeroGPT. `https://www.zerogpt.com/`, 2023.

X. Zhao, P. V. Ananth, L. Li, and Y.-X. Wang. Provable robust watermarking for AI-generated text. In *International Conference on Learning Representations*, 2024a. URL `https://openreview.net/forum?id=SsmT8aO45L`.

X. Zhao, L. Li, and Y.-X. Wang. Permute-and-Flip: An optimally robust and watermarkable decoder for llms. *arXiv preprint arXiv:2402.05864*, 2024b.