

What Can Statistics Offer to Language Models: Watermarking and Evaluation

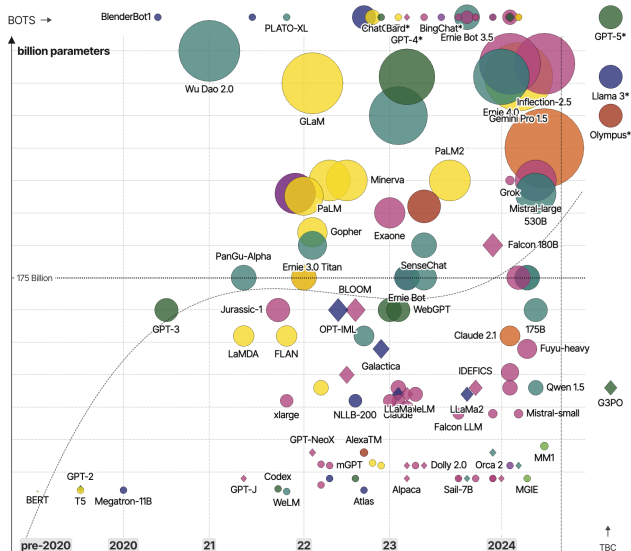
Xiang Li

University of Pennsylvania

Rutgers University

Jan. 30, 2025

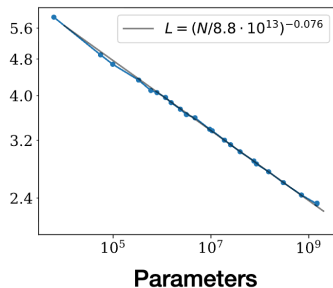
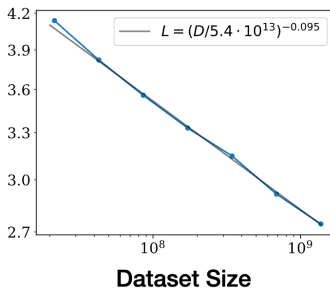
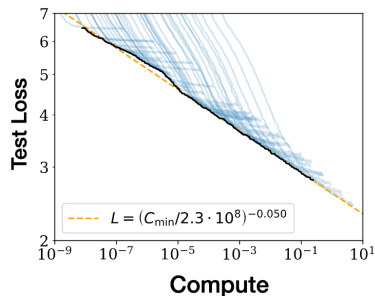
The rise of LLMs



David McCandless, Tom Evans, Paul Barton
Information is Beautiful // UPDATED 20th Mar 24

source: news reports, [LifeArchitect.ai](https://lifeaiarchitect.ai)
* = parameters undisclosed // see [the data](https://lifeaiarchitect.ai)

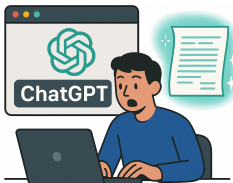
The scaling era [Kaplan et al., 2020]



Applications of LLMs

Previous

In this paper, we study a simple method for detecting distribution shift. The idea is to compare the empirical distribution of the generated tokens with the reference distribution and see if there is a significant difference. Although the method is easy to implement, the explanation of why it works is not very clear.



Polished

This paper investigates a straightforward approach for detecting distributional shifts by comparing the empirical token distribution with a reference model. Despite its computational simplicity, the method is supported by a clear statistical rationale: deviations in the empirical frequencies reveal discrepancies between the underlying generating processes.

Materia Polishing

Could you find any papers on the statistical foundations of large language models?

Pro · 2 steps · 20 sources

Statistical Foundations of Large Language Models: Key Research and Insights

A substantial body of research explores the statistical underpinnings of large language models (LLMs), tracing their evolution from early probabilistic models to modern deep learning architectures.

Core Statistical Principles in LLMs

LLMs are fundamentally statistical models that encode probability distributions over sequences of words, aiming to approximate the distribution found in large text corpora. Early models, such as n -gram models, estimated the probability of a word given its preceding context, but suffered from data sparsity and limited context length. Modern LLMs, built on neural architectures like Transformers, extend these ideas by learning complex, high-dimensional probability distributions and leveraging large-scale data and model scaling laws to improve performance [1][2][10].

Ask a follow up...

All Pro

Filter

Literature Review



Coding & Debugging

Attachment: [2503.10138v1.pdf]

Can you improve the condition on the step-size in Theorem 1? I don't want to add any more hypotheses, I just want you to work to improve the step-size condition under the same assumptions as Theorem 1.

[Reasoned for 17m 35s]

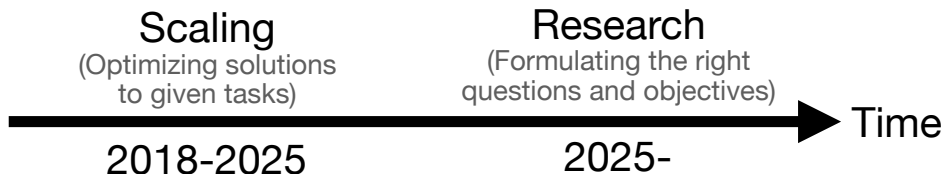
Yes. Under the same assumptions as Theorem 1 (convex L -smooth f and fixed step size for gradient descent), you can enlarge the guaranteed-convexity window from $\eta \leq 1/L$ to

$$\eta \leq \frac{3}{2L}$$

i.e., the optimization curve $n \mapsto f(x_n)$ is convex for any step size $\eta \in (0, 3/(2L)]$. This strictly improves Theorem 1's bound and still lies below the non-convexity regime exhibited in Theorem 2. See Theorem 1 and Theorem 2 (pp. 4-6) for the problem setup and statements referenced below.

Theorems Proofs

From scaling to research

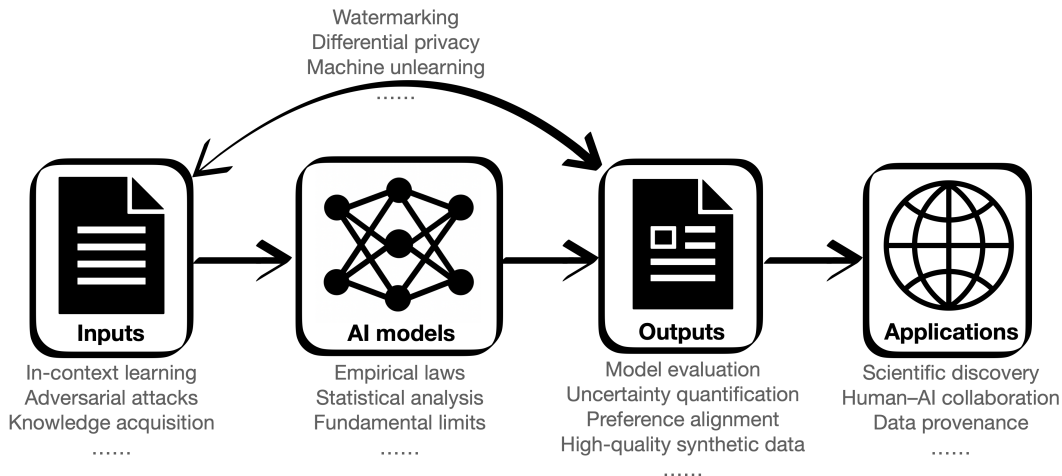


- *"Back to the age of research with big computers."* — Ilya Sutskever, 2025.
- Recent invited talks, such as Shunyu Yao's ICLR 2025 lecture, emphasize that future progress depends on formulating the right research questions.

What statistics can contribute to LLM research [Su, 2025]

- **A second-principle lens:** treat LLMs as complex, data-driven stochastic systems and study *quantitative regularities* of specific behaviors, without studying every mechanism.
- **Emergent simplifications at scale:** as system complexity grows, certain patterns become stable and analyzable.
- **Why it is practical now:** statistical approaches can provide *compute-light* first steps toward reliability—robustness, detection, evaluation, and uncertainty quantification.

Statistical foundation for LLMs



My work on statistical foundation of LLMs

- Language watermarking
 - Optimal detection [Li et al., 2025b, Cai et al., 2025]
 - Robust inference [Li et al., 2025a]
 - Proportion estimation [Li et al., 2025d]
 - Empirical study of GoF tests [He et al., 2025a]
 - Watermarking tabular data [Zhao et al., 2025b]
 - Inference acceleration [He et al., 2025b]
- Model evaluation [Li et al., 2025e]
- Uncertainty quantification [Xin et al., 2025]

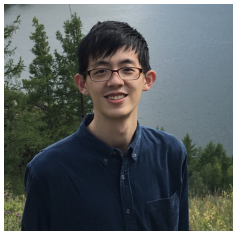
Part I: Optimal Detection for Language Watermarking

A Statistical Framework of Watermarks for Large Language Models: Pivot, Detection Efficiency and Optimal Rules

Annals of Statistics (2025) — Selected for JSM 2025 AoS Invited Paper Session

Robust Detection of Watermarks in Large Language Models under Human Edits

JRSS Series B (2025) — Selected for RSS 2026 Editors Invited Session



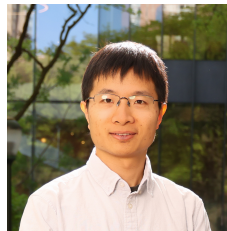
Feng Ruan



Huiyuan Wang



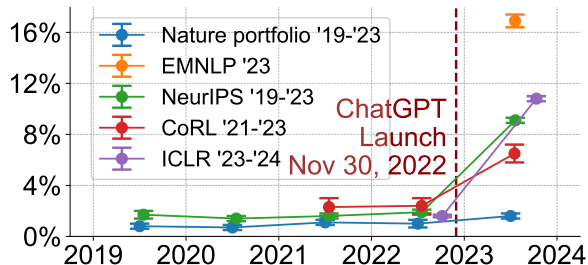
Qi Long



Weijie Su

Peer review or LLM-assisted review?

- Liang et al. [2024] finds that between 6.5% and 16.9% reviews of some ML conferences were substantially modified by LLMs.



Peer review or LLM review?

- ICLR 2026 is reportedly seeing a surge of LLM-generated reviews with extreme scores.
- Pangram Labs estimate that of 75,800 paper reviews, 21% were fully AI-generated and another 35% were AI-edited to varying degrees.
- Is your paper review really a peer's opinion or LLM's hallucination?



ICLR 2026
@iclr_conf

X.com

We are aware of low-quality and LLM-generated reviews and are currently deliberating on appropriate courses of action. For now, authors who receive very poor quality or LLM-generated reviews should flag them to their ACs. We appreciate the community's efforts in reporting these!

18:44 · 2025/11/16 · **5.7K** Views



Even use LLMs to write papers!

ARTICLE INFO

Keywords:

Lithium metal battery
Lithium dendrites
CuMOF-ANFs separator

ABSTRACT

Lithium metal, due to its advanced potential, is used as a negative of energy storage systems. However, poor safety, so lithium dendrites, the larger specific surface area (CuMOF-ANFs) composite separator, the discharge capacity. Li-Li batteries can continuously show that CuMOF-ANFs composite separator provides a new perspective.

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



Micah Goldblum
@micahgoldblum

X.com

An LLM-generated paper is in the top 17% of ICLR submissions in terms of average reviewer score, having received two 8's. The paper has tons of BS jargon and hallucinated references. Fortunately, one reviewer actually looked at the paper and gave it a zero. 1/3

16:50 · 2025/11/13 · **408K** Views



38



184



1.3K



340



How to detect AI-generated text (provably)?

- Use linguistic pattern differences [Mitchell et al., 2023, Bao et al., 2023, Yang et al., 2023]
- Train binary classifiers [GPTZero, 2023, ZeroGPT, 2023, Emi and Spero, 2024]

How to detect AI-generated text (provably)?

- Use linguistic pattern differences [Mitchell et al., 2023, Bao et al., 2023, Yang et al., 2023]
- Train binary classifiers [GPTZero, 2023, ZeroGPT, 2023, Emi and Spero, 2024]
- Empirically, inaccurate, unreliable [Weber-Wulff et al., 2023] & often biased [Krishna et al., 2024, Sadasivan et al., 2023, Liang et al., 2023]

It seems hopeless...

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



A principled approach: Watermarking LLM-generated text

Hope: LLMs are probabilistic machines, and we control how they generate texts.

A principled approach: Watermarking LLM-generated text

Hope: LLMs are probabilistic machines, and we control how they generate texts.

A watermark embeds a private, detectable signal into LLM-generated text by modifying the model's generation only [Kirchenbauer et al., 2023].

A principled approach: Watermarking LLM-generated text

Hope: LLMs are probabilistic machines, and we control how they generate texts.

A watermark embeds a private, detectable signal into LLM-generated text by modifying the model's generation only [Kirchenbauer et al., 2023].

- Creates a statistical dependency between the visible text and a private information.

A principled approach: Watermarking LLM-generated text

Hope: LLMs are probabilistic machines, and we control how they generate texts.

A watermark embeds a private, detectable signal into LLM-generated text by modifying the model's generation only [Kirchenbauer et al., 2023].

- Creates a statistical dependency between the visible text and a private information.
- Unlikely to appear in human-written text.

A principled approach: Watermarking LLM-generated text

Hope: LLMs are probabilistic machines, and we control how they generate texts.

A watermark embeds a private, detectable signal into LLM-generated text by modifying the model's generation only [Kirchenbauer et al., 2023].

- Creates a statistical dependency between the visible text and a private information.
- Unlikely to appear in human-written text.
- Efficient, provable, and slight (or no) degradation on text quality.

An active research area with practical importance

A Zoo of Watermarking Schemes (since Jan 2023):

[Kirchenbauer et al., 2023, Aaronson, 2023, Kuditipudi et al., 2024, Zhao et al., 2025a, Christ et al., 2024, Wu et al., 2024, Hu et al., 2024, Kirchenbauer et al., 2024, Zhao et al., 2024, Xie et al., 2025b, Fu et al., 2024, Dathathri et al., 2024, Gloaguen et al., 2025, Abdalla and Vershynin, 2025].

An active research area with practical importance

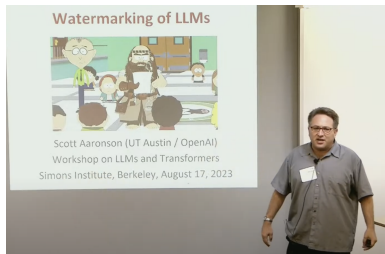
A Zoo of Watermarking Schemes (since Jan 2023):

[Kirchenbauer et al., 2023, Aaronson, 2023, Kuditipudi et al., 2024, Zhao et al., 2025a, Christ et al., 2024, Wu et al., 2024, Hu et al., 2024, Kirchenbauer et al., 2024, Zhao et al., 2024, Xie et al., 2025b, Fu et al., 2024, Dathathri et al., 2024, Gloaguen et al., 2025, Abdalla and Vershynin, 2025].

- **Open-source toolkits** have been developed to support research [Pan et al., 2024].
- **Large-scale empirical studies** benchmark watermarking methods [Fernandez et al., 2023, Gloaguen et al., 2025, Piet et al., 2023, He et al., 2025a].
- **Applications:** Academic integrity [Xie et al., 2025a], Biosecurity for protein design [Chen et al., 2025], Misinformation control [Shumailov et al., 2023], Content transparency, and Attribution & Liability.

Industry commitment

In July 2023, several AI giants like OpenAI, Anthropic, Google, and Meta pledged to the White House that they will voluntarily research and develop watermarking options.



Gumbel-max (OpenAI)

Article | [Open access](#) | Published: 23 October 2024

Scalable watermarking for identifying large language model outputs

[Sumanth Dathathri](#) , [Abigail See](#), [Sumedh Ghaisas](#), [Po-Sen Huang](#), [Rob McAdam](#), [Johannes Welbl](#), [Vandana Bachani](#), [Alex Kaskasoli](#), [Robert Stanforth](#), [Tatiana Matejovicova](#), [Jamie Hayes](#), [Nidhi Vyas](#), [Majd Al Merey](#), [Jonah Brown-Cohen](#), [Rudy Bunel](#), [Borja Balle](#), [Taylan Cemgil](#), [Zahra Ahmed](#), [Kitty Stacpoole](#), [Ilia Shumailov](#), [Ciprian Baetu](#), [Sven Gowal](#), [Demis Hassabis](#) & [Pushmeet Kohli](#) 

[Nature](#) **634**, 818–823 (2024) | [Cite this article](#)

115k Accesses | **15** Citations | **986** Altmetric | [Metrics](#)

SynthID (Google Deepmind)

Statistical questions for LLM watermarks

Two types of errors

- **Type I error** – falsely flagging human prose as AI-generated
- **Type II error** – failing to detect AI-generated text

Evaluation of watermarks

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

Tokens: Smallest units of LLM generation

- LLMs generate text by gathering many small units, called “tokens”.
- Tokens can be words, parts of words, or even punctuation marks.

Tokens: Smallest units of LLM generation

- LLMs generate text by gathering many small units, called “tokens”.
- Tokens can be words, parts of words, or even punctuation marks.

GPT-5.x & O1/3

GPT-4 & GPT-3.5 (legacy)

GPT-3 (legacy)

Rutgers University is a top-tier public research university with an exceptionally strong academic and research reputation.

Rutgers University is a top-tier public research university with an exceptionally strong academic and research reputation.

[194209, 13027, 4923, 382, 261, 2344, 98409, 751, 4176, 16490, 483, 448, 62431, 5532, 15487, 326, 4176, 22785, 558]

<https://platform.openai.com/tokenizer>

Autoregressive generation: How LLMs combine tokens

- The vocabulary $\mathcal{W} = \{1, \dots, K\}$, a token w_t , and a text $w_{1:(t-1)} := w_1 \cdots w_{t-1}$.

Autoregressive generation: How LLMs combine tokens

- The vocabulary $\mathcal{W} = \{1, \dots, K\}$, a token w_t , and a text $w_{1:(t-1)} := w_1 \cdots w_{t-1}$.
- **Autoregressiveness**: An LLM generates each token sequentially by sampling from a probability distribution conditioned on previous tokens:

$w_t \sim \mathbf{P}_t$ where $\mathbf{P}_t = \text{LLM}(w_{1:(t-1)})$ is next-token prediction (NTP) dist. on $\mathcal{W}$.

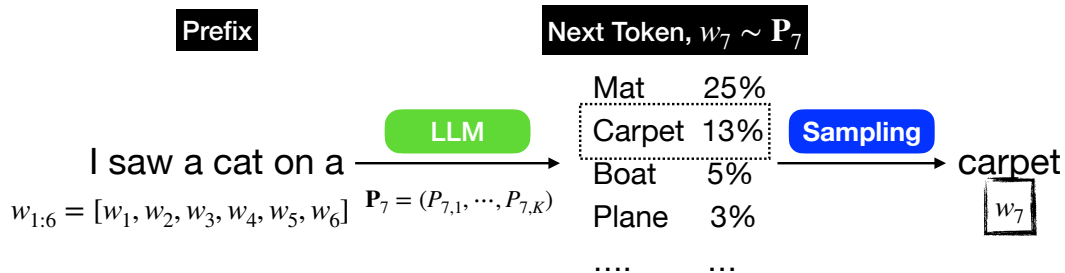
Autoregressive generation: How LLMs combine tokens

- The vocabulary $\mathcal{W} = \{1, \dots, K\}$, a token w_t , and a text $w_{1:(t-1)} := w_1 \cdots w_{t-1}$.
- **Autoregressiveness:** An LLM generates each token sequentially by sampling from a probability distribution conditioned on previous tokens:

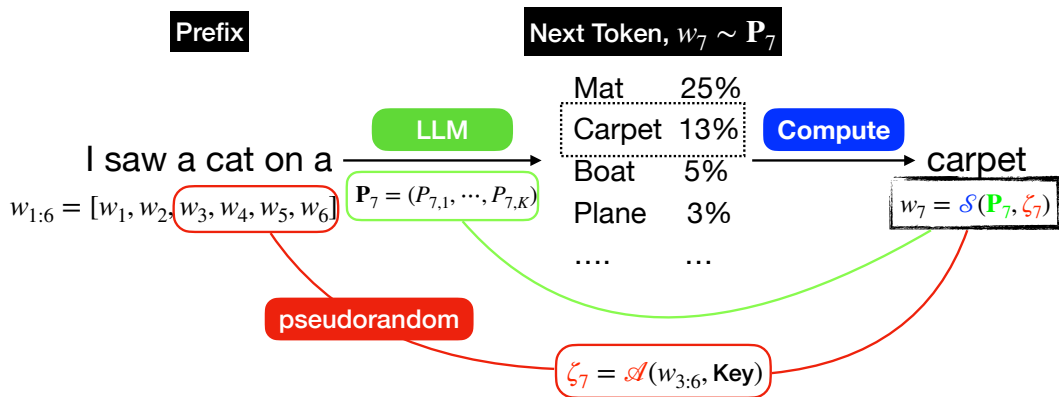
$w_t \sim \mathbf{P}_t$ where $\mathbf{P}_t = \text{LLM}(w_{1:(t-1)})$ is next-token prediction (NTP) dist. on $\mathcal{W}$.

- **Practical constraint:** $\mathbf{P}_t$ is unknown due to (i) limited access to LLMs and (ii) unknown system/user prompts.

Autoregressive generation: An illustration

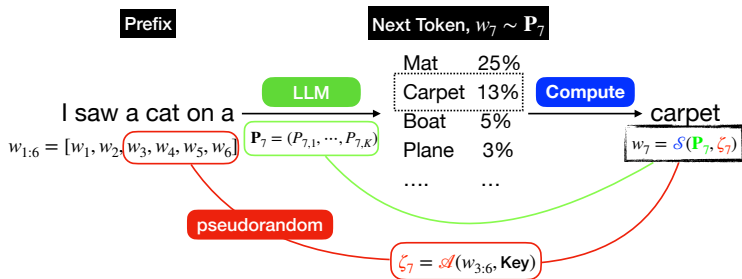


Autoregressive generation with watermark: An illustration



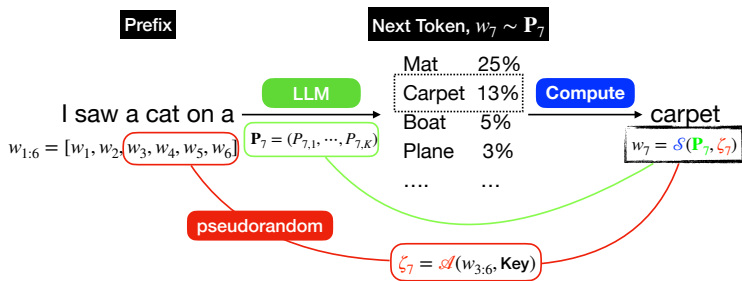
Autoregressive generation with watermark: An illustration

- Reparameterization trick.



Autoregressive generation with watermark: An illustration

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

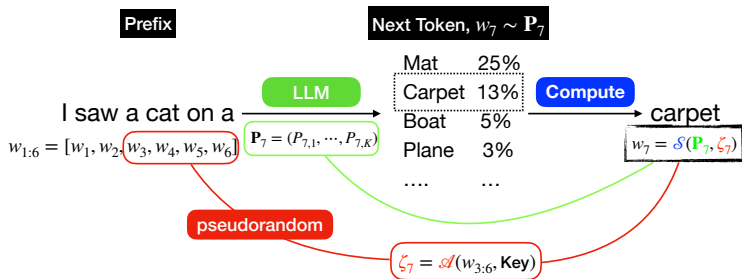


Autoregressive generation with watermark: An illustration

- Reparameterization trick.
- $\mathcal{A}$ is a hash function and $\mathcal{S}(\mathbf{P}, \zeta)$ is a (deterministic) decoder.
- Unbiased: for any token w ,

$$\mathbb{P}_{\zeta}(\mathcal{S}(\mathbf{P}, \zeta) = w) = P_w.$$

Does not degrade text quality.



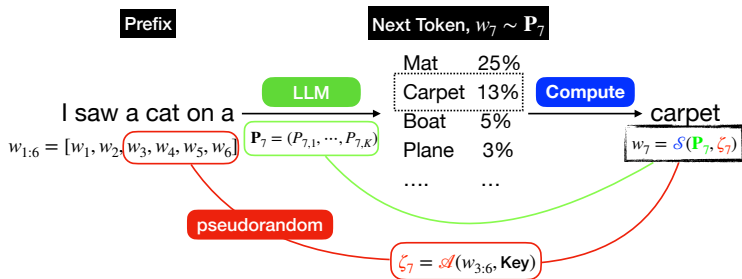
Autoregressive generation with watermark: An illustration

- Reparameterization trick.
- $\mathcal{A}$ is a hash function and $\mathcal{S}(\mathbf{P}, \zeta)$ is a (deterministic) decoder.
- Unbiased: for any token w ,

$$\mathbb{P}_{\zeta}(\mathcal{S}(\mathbf{P}, \zeta) = w) = P_w.$$

Does not degrade text quality.

- The watermark signal is the statistical dependence between w_t and ζ_t .



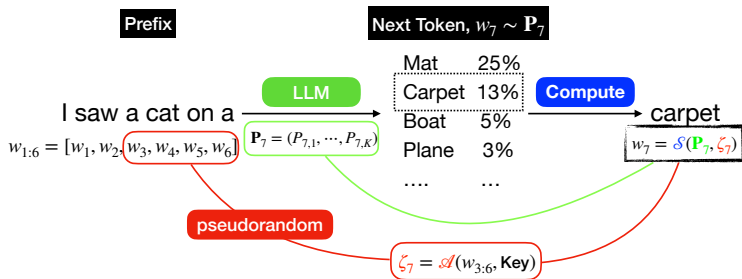
Autoregressive generation with watermark: An illustration

- Reparameterization trick.
- $\mathcal{A}$ is a hash function and $\mathcal{S}(\mathbf{P}, \zeta)$ is a (deterministic) decoder.
- Unbiased: for any token w ,

$$\mathbb{P}_{\zeta}(\mathcal{S}(\mathbf{P}, \zeta) = w) = P_w.$$

Does not degrade text quality.

- The watermark signal is the statistical dependence between w_t and ζ_t .
- $\zeta_{1:n}$ are i.i.d. but recoverable.



Comments on pseudorandom numbers

$\mathcal{A}$ (public data, private info) computes the pseudorandom number:

- Public data = prior tokens $w_{1:(t-1)}$ or a segment $w_{(t-m):t}$ [Kirchenbauer et al., 2023].
- Private info = secret key, denoted Key.
- Well-developed in theoretical computer science [Barak, 2021].
 - Pseudorandom numbers are “reproducible” **iff** Key is known.
 - Key will be shared with the verifier through a secure protocol.

Comments on pseudorandom numbers

$\mathcal{A}$ (public data, private info) computes the pseudorandom number:

- Public data = prior tokens $w_{1:(t-1)}$ or a segment $w_{(t-m):t}$ [Kirchenbauer et al., 2023].
- Private info = secret key, denoted Key.
- Well-developed in theoretical computer science [Barak, 2021].
 - Pseudorandom numbers are “reproducible” **iff** Key is known.
 - Key will be shared with the verifier through a secure protocol.

$\mathcal{A}$ is statistically sensitive: outputs of different inputs will be independent.

Property (*Soundness of pseudorandomness*)

1. $\zeta_t = \mathcal{A}(w_{1:t-1}, \text{Key})$ for $t = 1, \dots, n$ are iid copies of a known random variable ζ .
2. ζ_t is statistically independent of $w_{1:t-1}$.

Comments on pseudorandom numbers

$\mathcal{A}$ (public data, private info) computes the pseudorandom number:

- Public data = prior tokens $w_{1:(t-1)}$ or a segment $w_{(t-m):t}$ [Kirchenbauer et al., 2023].
- Private info = secret key, denoted Key.
- Well-developed in theoretical computer science [Barak, 2021].
 - Pseudorandom numbers are “reproducible” **iff** Key is known.
 - Key will be shared with the verifier through a secure protocol.

$\mathcal{A}$ is statistically sensitive: outputs of different inputs will be independent.

Property (*Soundness of pseudorandomness*)

1. $\zeta_t = \mathcal{A}(w_{1:t-1}, \text{Key})$ for $t = 1, \dots, n$ are iid copies of a known random variable ζ .
2. ζ_t is statistically independent of $w_{1:t-1}$.

- Theory assumes true randomness; Implementation use deterministic generation.
- Similar to setting a *seed* for reproducibility in simulations.

The simplest unbiased watermark: Baby watermark

Definition (*Unbiasedness*)

A watermark is unbiased if the marginal distribution of w in (w, ζ) is still $\mathbf{P}$, i.e.,

$$\mathbb{P}_{\zeta}(\mathcal{S}(\mathbf{P}, \zeta) = w) = P_w \text{ for any } \mathbf{P} \text{ and } w \in \mathcal{W}.$$

The simplest unbiased watermark: Baby watermark

Definition (*Unbiasedness*)

A watermark is unbiased if the marginal distribution of w in (w, ζ) is still $\mathbf{P}$, i.e.,

$$\mathbb{P}_{\zeta}(\mathcal{S}(\mathbf{P}, \zeta) = w) = P_w \text{ for any } \mathbf{P} \text{ and } w \in \mathcal{W}.$$

- Let $\mathcal{W} = \{0, 1\}$, $\mathbf{P}_t = (P_{t,0}, P_{t,1})$, and ζ_t be i.i.d. copies of $\mathcal{U}(0, 1)$.
- Decoder:

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

The simplest unbiased watermark: Baby watermark

Definition (*Unbiasedness*)

A watermark is unbiased if the marginal distribution of w in (w, ζ) is still $\mathbf{P}$, i.e.,

$$\mathbb{P}_{\zeta}(\mathcal{S}(\mathbf{P}, \zeta) = w) = P_w \text{ for any } \mathbf{P} \text{ and } w \in \mathcal{W}.$$

- Let $\mathcal{W} = \{0, 1\}$, $\mathbf{P}_t = (P_{t,0}, P_{t,1})$, and ζ_t be i.i.d. copies of $\mathcal{U}(0, 1)$.
- Decoder:

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

- It is unbiased. If ζ_t is large, w_t is more likely to be 1.
- Statistic score for detection: $\sum_{t=1}^n (2w_t - 1)(2\zeta_t - 1)$.

The simplest unbiased watermark: Baby watermark

Definition (*Unbiasedness*)

A watermark is unbiased if the marginal distribution of w in (w, ζ) is still $\mathbf{P}$, i.e.,

$$\mathbb{P}_{\zeta}(\mathcal{S}(\mathbf{P}, \zeta) = w) = P_w \text{ for any } \mathbf{P} \text{ and } w \in \mathcal{W}.$$

- Let $\mathcal{W} = \{0, 1\}$, $\mathbf{P}_t = (P_{t,0}, P_{t,1})$, and ζ_t be i.i.d. copies of $\mathcal{U}(0, 1)$.
- Decoder:

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

- It is unbiased. If ζ_t is large, w_t is more likely to be 1.
- Statistic score for detection: $\sum_{t=1}^n (2w_t - 1)(2\zeta_t - 1)$.

Take-away

Statistically, an unbiased watermark = a sampling method from each $\mathbf{P}_t$.

Our focus: Gumbel-max [Aaronson, 2023]

Definition (*Gumbel-max trick* [Gumbel, 1948])

Let $U_1, \dots, U_K$ be i.i.d. copies of $\mathcal{U}(0, 1)$. Then,

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

Our focus: Gumbel-max [Aaronson, 2023]

Definition (*Gumbel-max trick* [Gumbel, 1948])

Let $U_1, \dots, U_K$ be i.i.d. copies of $\mathcal{U}(0, 1)$. Then,

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

Definition (*Gumbel-max watermark*)

With $\zeta_t = (U_1, \dots, U_K) = \mathcal{A}(w_{1:(t-1)}, \text{Key})$ (often depending on the last 5 tokens),

$$w_t = \mathcal{S}^{\text{gum}}(\mathbf{P}_t, \zeta_t) = \arg \max_{w \in \mathcal{W}} \frac{\log U_{t,w}}{P_{t,w}}.$$

- Implemented internally at OpenAI.

Our focus: Gumbel-max [Aaronson, 2023]

Definition (*Gumbel-max trick* [Gumbel, 1948])

Let $U_1, \dots, U_K$ be i.i.d. copies of $\mathcal{U}(0, 1)$. Then,

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

Definition (*Gumbel-max watermark*)

With $\zeta_t = (U_1, \dots, U_K) = \mathcal{A}(w_{1:(t-1)}, \text{Key})$ (often depending on the last 5 tokens),

$$w_t = \mathcal{S}^{\text{gum}}(\mathbf{P}_t, \zeta_t) = \arg \max_{w \in \mathcal{W}} \frac{\log U_{t,w}}{P_{t,w}}.$$

- Implemented internally at OpenAI.
- Unbiased as $w_t \sim \mathbf{P}_t$.

Our focus: Gumbel-max [Aaronson, 2023]

Definition (*Gumbel-max trick* [Gumbel, 1948])

Let $U_1, \dots, U_K$ be i.i.d. copies of $\mathcal{U}(0, 1)$. Then,

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

Definition (*Gumbel-max watermark*)

With $\zeta_t = (U_1, \dots, U_K) = \mathcal{A}(w_{1:(t-1)}, \text{Key})$ (often depending on the last 5 tokens),

$$w_t = \mathcal{S}^{\text{gum}}(\mathbf{P}_t, \zeta_t) = \arg \max_{w \in \mathcal{W}} \frac{\log U_{t,w}}{P_{t,w}}.$$

- Implemented internally at OpenAI.
- Unbiased as $w_t \sim \mathbf{P}_t$.
- Embedded signal: selected U_{w_t} tends to be larger.

Hypothesis testing formulation

- Can always compute $\zeta_t = \mathcal{A}(w_{1:(t-1)}, \text{Key})$.
- Dataset = tokens $w_{1:n} := w_1 w_2 \cdots w_n$ + pseudorandom $\zeta_{1:n} := \zeta_1 \zeta_2 \cdots \zeta_n$.

Human-written

When text is written by a human, w_t, ζ_t are independent, since the human simply cannot compute ζ_t .

LLM-generated

When text is generated by the LLM, w_t, ζ_t are dependent because of the decoder:
 $w_t = \mathcal{S}(\mathbf{P}_t, \zeta_t)$.

Hypothesis testing formulation

- Can always compute $\zeta_t = \mathcal{A}(w_{1:(t-1)}, \text{Key})$.
- Dataset = tokens $w_{1:n} := w_1 w_2 \cdots w_n$ + pseudorandom $\zeta_{1:n} := \zeta_1 \zeta_2 \cdots \zeta_n$.

Human-written

When text is written by a human, w_t, ζ_t are independent, since the human simply cannot compute ζ_t .

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

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

LLM-generated

When text is generated by the LLM, w_t, ζ_t are dependent because of the decoder:
 $w_t = \mathcal{S}(\mathbf{P}_t, \zeta_t)$.

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

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

A challenge: Unknown NTP distributions

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

Hypothesis testing

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

A challenge: Unknown NTP distributions

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

Hypothesis testing

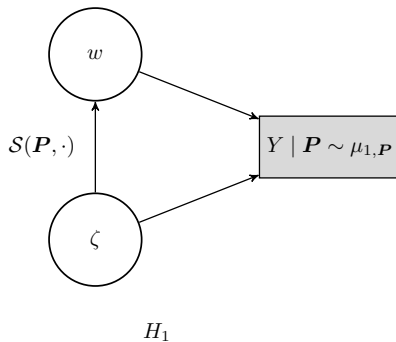
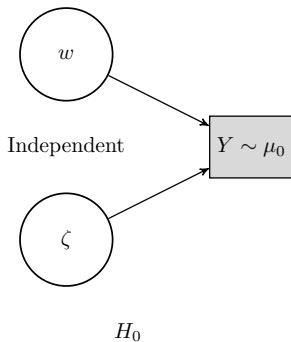
- Under H_0 , $(w_t, \zeta_t) \mid (w_{1:(t-1)}, \zeta_{1:(t-1)}) \stackrel{d}{=} \mathbf{P}_t \times \zeta$
- Under H_1 , $(w_t, \zeta_t) \mid (w_{1:(t-1)}, \zeta_{1:(t-1)}) \stackrel{d}{=} (\mathcal{S}(\zeta, \mathbf{P}_t), \zeta)$
- Conservative control of Type I errors due to unknown $\mathbf{P}_t$'s.
- Neyman-Pearson lemma resorts to the log-likelihood ratio test:

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

A practical approach: Pivot under the null [Li et al., 2025b]

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

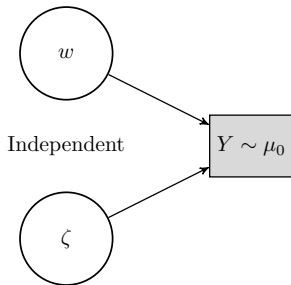
- Under H_0 , $Y_t \sim \mu_0$, regardless of $\mathbf{P}_t$
- Under H_1 , $Y_t \sim Y(\mathcal{S}(\zeta_t, \mathbf{P}_t), \zeta_t) \stackrel{d}{=} \mu_{1, \mathbf{P}_t}$



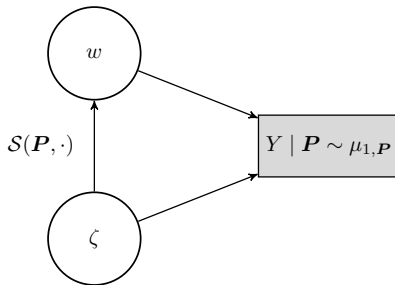
A practical approach: Pivot under the null [Li et al., 2025b]

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

- Under H_0 , $Y_t \sim \mu_0$, regardless of $\mathbf{P}_t$
- Under H_1 , $Y_t \sim Y(\mathcal{S}(\zeta_t, \mathbf{P}_t), \zeta_t) \stackrel{d}{=} \mu_{1, \mathbf{P}_t}$



H_0



H_1

Example: $Y_t = (2w_t - 1)(2\zeta_t - 1) \sim \mathcal{U}(-1, 1)$ under H_0 for the baby watermark.

Hypothesis testing via pivoting

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

- Under H_0 , $Y_t \sim \mu_0$, regardless of $\mathbf{P}_t$
- Under H_1 , $Y_t \sim Y(\mathcal{S}(\zeta_t, \mathbf{P}_t), \zeta_t) \stackrel{d}{=} \mu_{1, \mathbf{P}_t}$

Hypothesis testing via pivoting

$$H_0 : Y_t \stackrel{iid}{\sim} \mu_0, \quad t = 1, \dots, n \quad \text{vs} \quad H_1 : Y_t | \mathbf{P}_t \sim \mu_{1, \mathbf{P}_t}, \quad t = 1, \dots, n$$

- Test distributional difference rather than independence.

Hypothesis testing via pivoting

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

- Under H_0 , $Y_t \sim \mu_0$, regardless of $\mathbf{P}_t$
- Under H_1 , $Y_t \sim Y(\mathcal{S}(\zeta_t, \mathbf{P}_t), \zeta_t) \stackrel{d}{=} \mu_{1, \mathbf{P}_t}$

Hypothesis testing via pivoting

$$H_0 : Y_t \stackrel{iid}{\sim} \mu_0, \quad t = 1, \dots, n \quad \text{vs} \quad H_1 : Y_t | \mathbf{P}_t \sim \mu_{1, \mathbf{P}_t}, \quad t = 1, \dots, n$$

- Test distributional difference rather than independence.
- Tight Type I error control.

Hypothesis testing via pivoting

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

- Under H_0 , $Y_t \sim \mu_0$, regardless of $\mathbf{P}_t$
- Under H_1 , $Y_t \sim Y(\mathcal{S}(\zeta_t, \mathbf{P}_t), \zeta_t) \stackrel{d}{=} \mu_{1, \mathbf{P}_t}$

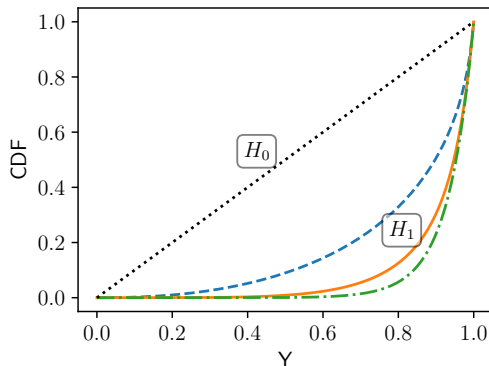
Hypothesis testing via pivoting

$$H_0 : Y_t \stackrel{iid}{\sim} \mu_0, \quad t = 1, \dots, n \quad \text{vs} \quad H_1 : Y_t | \mathbf{P}_t \sim \mu_{1, \mathbf{P}_t}, \quad t = 1, \dots, n$$

- Test distributional difference rather than independence.
- Tight Type I error control.
- Not unique, may lead to information loss, but convenient.
- Test score $T_h = \sum_{t=1}^n h(Y_t)$ for some score function h . Reject H_0 if T_h is larger than a threshold.

Pivot for Gumbel-max watermark

- Pivotal statistic: $Y_t^{\text{ars}} = U_{t,w_t}$ where $\zeta_t = (U_{t,1}, \dots, U_{t,K})$ with w_t selected.
- Under H_0 , $Y_t^{\text{ars}} \stackrel{iid}{\sim} \mu_0 = \mathcal{U}(0, 1)$.
- Under H_1 , the CDF of $\mu_{1,\mathbf{P}_t}$ is $\mathbb{P}_1(Y_t^{\text{ars}} \leq r | \mathbf{P}_t) = \sum_{k=1}^K P_{t,k} r^{1/P_{t,k}}$.



Detection for Gumbel-max watermark

Definition (*Default detection for Gumbel-max*)

Aaronson [2023] rejects H_0 if the following $T_{h_{\text{ars}}}$ is larger than a given threshold:

$$T_{h_{\text{ars}}} = \sum_{t=1}^n h_{\text{ars}}(Y_t^{\text{ars}}) \text{ where } h_{\text{ars}}(y) = -\log(1 - y).$$

- Under H_0 , $\mathbb{E}_0[T_{\text{ars}}] = n$ as $h_{\text{ars}}(Y_t^{\text{ars}}) \stackrel{iid}{\sim} \text{Exp}(1)$.
- Under H_1 , $\mathbb{E}_1[T_{\text{ars}}] \geq n + \left(\frac{\pi^2}{6} - 1\right) \sum_{t=1}^n \mathbb{E}_1 \text{Ent}(\mathbf{P}_t)$ where $\text{Ent}(\mathbf{P}_t)$ is Shannon entropy $-\sum_{k=1}^K P_{t,k} \log P_{t,k}$.

Detection for Gumbel-max watermark

Definition (*Default detection for Gumbel-max*)

Aaronson [2023] rejects H_0 if the following $T_{h_{\text{ars}}}$ is larger than a given threshold:

$$T_{h_{\text{ars}}} = \sum_{t=1}^n h_{\text{ars}}(Y_t^{\text{ars}}) \text{ where } h_{\text{ars}}(y) = -\log(1 - y).$$

- Under H_0 , $\mathbb{E}_0[T_{\text{ars}}] = n$ as $h_{\text{ars}}(Y_t^{\text{ars}}) \stackrel{iid}{\sim} \text{Exp}(1)$.
- Under H_1 , $\mathbb{E}_1[T_{\text{ars}}] \geq n + \left(\frac{\pi^2}{6} - 1\right) \sum_{t=1}^n \mathbb{E}_1 \text{Ent}(\mathbf{P}_t)$ where $\text{Ent}(\mathbf{P}_t)$ is Shannon entropy $-\sum_{k=1}^K P_{t,k} \log P_{t,k}$.
- Fernandez et al. [2023] proposes $\log y$ and finds $-\log(1 - y)$ works better than it.

From a statistical perspective

Questions

- How to theoretically compare different score functions?
- What is the “optimal” score function?

Detection efficiency

Fixing Type I error, a detection rule is preferred if it has a higher power.

- However, the comparison depends on the unknown NTP distributions $P_{1:n}$.

Detection efficiency

Fixing Type I error, a detection rule is preferred if it has a higher power.

- However, the comparison depends on the unknown NTP distributions $\mathbf{P}_{1:n}$.

Minimax viewpoint: the lowest power over all NTP distributions?

- All detection rules are powerless as $Y(w_t, \zeta_t)$ has the same distribution under H_0 and H_1 if $\mathbf{P}_t$ has an entry of 1.

Detection efficiency

Fixing Type I error, a detection rule is preferred if it has a higher power.

- However, the comparison depends on the unknown NTP distributions $\mathbf{P}_{1:n}$.

Minimax viewpoint: the lowest power over all NTP distributions?

- All detection rules are powerless as $Y(w_t, \zeta_t)$ has the same distribution under H_0 and H_1 if $\mathbf{P}_t$ has an entry of 1.

Class-dependent efficiency

- Select a class $\mathcal{P}$ that is believed to contain all $\mathbf{P}_{1:n}$.
- Evaluate efficiency by the least-favorable power attained over $\mathcal{P}$.
- What is a reasonable class $\mathcal{P}$?

Detection efficiency

Fixing Type I error, a detection rule is preferred if it has a higher power.

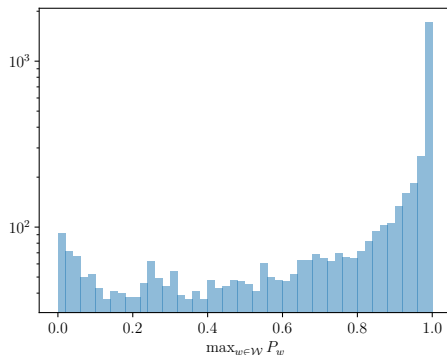
- However, the comparison depends on the unknown NTP distributions $\mathbf{P}_{1:n}$.

Minimax viewpoint: the lowest power over all NTP distributions?

- All detection rules are powerless as $Y(w_t, \zeta_t)$ has the same distribution under H_0 and H_1 if $\mathbf{P}_t$ has an entry of 1.

Class-dependent efficiency

- Select a class $\mathcal{P}$ that is believed to contain all $\mathbf{P}_{1:n}$.
- Evaluate efficiency by the least-favorable power attained over $\mathcal{P}$.
- What is a reasonable class $\mathcal{P}$?



Top prob. from ChatGPT-3.5-turbo.

Detection efficiency

Fixing Type I error, a detection rule is preferred if it has a higher power.

- However, the comparison depends on the unknown NTP distributions $\mathbf{P}_{1:n}$.

Minimax viewpoint: the lowest power over all NTP distributions?

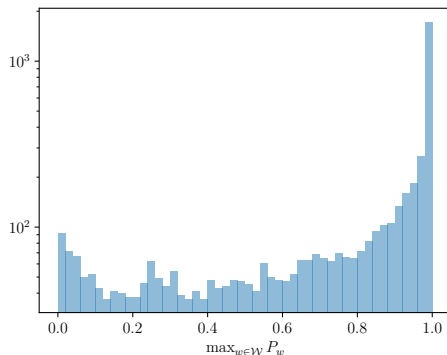
- All detection rules are powerless as $Y(w_t, \zeta_t)$ has the same distribution under H_0 and H_1 if $\mathbf{P}_t$ has an entry of 1.

Class-dependent efficiency

- Select a class $\mathcal{P}$ that is believed to contain all $\mathbf{P}_{1:n}$.
- Evaluate efficiency by the least-favorable power attained over $\mathcal{P}$.
- What is a reasonable class $\mathcal{P}$?

Δ -regular distribution class:

$$\mathcal{P}_\Delta := \{\mathbf{P} = (P_1, \dots, P_K) : \max_w P_w \leq 1 - \Delta\}.$$



Top prob. from ChatGPT-3.5-turbo.

Asymptotic class-dependent efficiency

Theorem (*Class-dependent efficiency*)

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

$$\limsup_{n \rightarrow \infty} \sup_{\text{all } \mathbf{P}_t \in \mathcal{P}} [\text{Type II error}]^{\frac{1}{n}} \leq \exp(-R_{\mathcal{P}}(h)),$$

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

$$R_{\mathcal{P}}(h) = - \inf_{\theta \geq 0} \left\{ \theta \mathbb{E}_0[h(Y)] + \sup_{\mathbf{P} \in \mathcal{P}} \mathbb{E}_{1, \mathbf{P}} \log \left([e^{-\theta h(Y)}] \right) \right\}.$$

Asymptotic class-dependent efficiency

Theorem (*Class-dependent efficiency*)

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

$$\limsup_{n \rightarrow \infty} \sup_{\text{all } \mathcal{P}_t \in \mathcal{P}} [\text{Type II error}]^{\frac{1}{n}} \leq \exp(-R_{\mathcal{P}}(h)),$$

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

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

- $R_{\mathcal{P}}(h)$ is also known as Cramér rate function.
- Tight in the minimax sense. Bahadur efficiency when $\mathcal{P}$ is a singleton.
- Monotonicity: $R_{\mathcal{P}_1}(h) \geq R_{\mathcal{P}_2}(h)$ if $\mathcal{P}_1 \subset \mathcal{P}_2$.
- $R_{\mathcal{P}}(h) = 0$ if $\mathcal{P}$ contains $(1, 0, \dots, 0)$, thereby justifying $\mathcal{P}_{\Delta}$.

Optimal score via class-dependent efficiency

Definition (*Optimal score function*)

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

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

- The minimax problem $\min_h \max_{\mathcal{P}} L(h, \mathbf{P})$ is generally not convex-concave.
- Case-by-case analysis is required, but we are often lucky.

How to maximize $R_{\mathcal{P}}(h)$

- Find the saddle point $(\mathbf{P}^*, h^*)$ that solves the minimax problem:

$$\min_h \max_{\mathbf{P} \in \mathcal{P}} L(h, \mathbf{P}) = \inf_h \left\{ \mathbb{E}_0[h(Y)] + \sup_{\mathbf{P} \in \mathcal{P}} \log \left(\mathbb{E}_1[e^{-h(Y)} | \mathbf{P}] \right) \right\}.$$

Theorem (*Saddle point condition*)

If there exists an $\mathbf{P}^* \in \mathcal{P}$ and a score function class $\mathcal{H}$ such that for all $h \in \mathcal{H}$,

$$h^* := \log \frac{d\mu_{1, \mathbf{P}^*}}{d\mu_0} \in \mathcal{H}, \quad (h^*)$$

$$\sup_{\mathbf{P} \in \mathcal{P}} \mathbb{E}_1[e^{-h(Y)} | \mathbf{P}] = \mathbb{E}_1[e^{-h(Y)} | \mathbf{P}^*], \quad (\mathbf{P}^*)$$

we then have

$$\max_h R_{\mathcal{P}}(h) = L(h^*, \mathbf{P}^*) = D_{\text{KL}}(\mu_0, \mu_{1, \mathbf{P}^*}),$$

where the maximum is obtained at h^* .

Optimal score for Gumbel-max watermark

Theorem (*Optimal score for Gumbel-max watermark*)

The optimal score that maximizes $h_{\Delta}^* := \arg \max_h R_{\mathcal{P}_{\Delta}}(h)$ is

$$h_{\text{gum},\Delta}(y) = \log \frac{d\mu_{1,\mathbf{P}_{\Delta}^*}}{d\mu_0}(y),$$

where

$$\mathbf{P}_{\Delta}^* = \left(\underbrace{1 - \Delta, \dots, 1 - \Delta}_{\lfloor \frac{1}{1-\Delta} \rfloor \text{ times}}, 1 - (1 - \Delta) \cdot \left\lfloor \frac{1}{1 - \Delta} \right\rfloor, 0, \dots \right).$$

Optimal score for Gumbel-max watermark

Theorem (*Optimal score for Gumbel-max watermark*)

The optimal score that maximizes $h_{\Delta}^* := \arg \max_h R_{\mathcal{P}_{\Delta}}(h)$ is

$$h_{\text{gum},\Delta}(y) = \log \frac{d\mu_{1,\mathbf{P}_{\Delta}^*}}{d\mu_0}(y),$$

where

$$\mathbf{P}_{\Delta}^* = \left(\underbrace{1 - \Delta, \dots, 1 - \Delta}_{\lfloor \frac{1}{1-\Delta} \rfloor \text{ times}}, 1 - (1 - \Delta) \cdot \left\lfloor \frac{1}{1 - \Delta} \right\rfloor, 0, \dots \right).$$

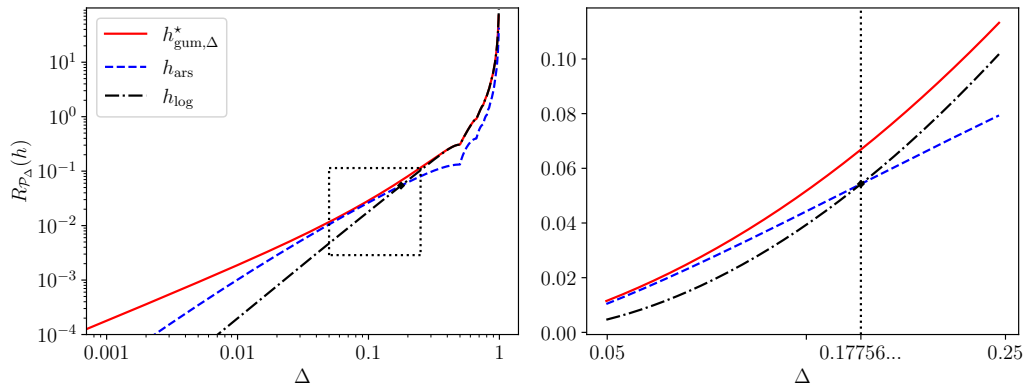
Key observation: $\mathbb{E}_1[e^{-h(Y)} | \mathbf{P}]$ is convex in $\mathbf{P}$ for any non-decreasing h .

($\mathbf{P}^*$) $\mathbf{P}_{\Delta}^*$ is the only extreme point in $\mathcal{P}_{\Delta}$ up to permutation.

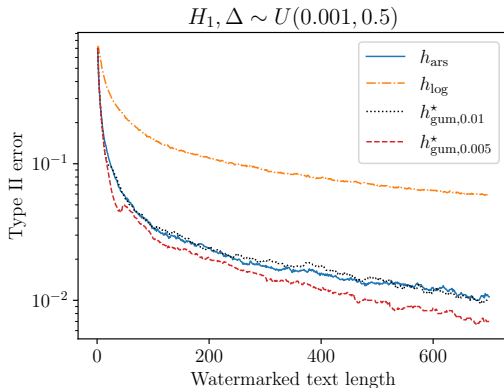
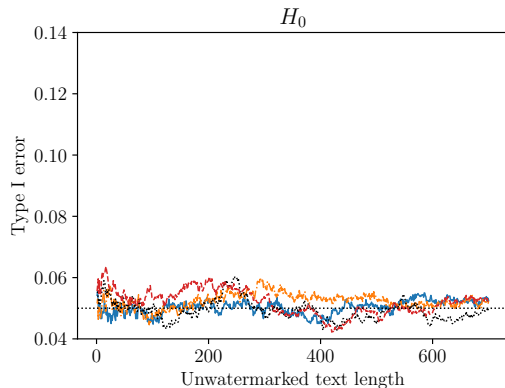
(h^*) When $0 < \Delta < 0.5$, $h_{\text{gum},\Delta} = \log(y^{\frac{\Delta}{1-\Delta}} + y^{\frac{1-\Delta}{\Delta}}) \uparrow$ as $\mathbf{P}_{\Delta}^* = (1 - \Delta, \Delta, 0, \dots)$.

Efficiency plot for Gumbel-max watermark

- Ars score: $h_{\text{ars}}(y) = -\log(1 - y)$ [Aaronson, 2023].
- Log score: $h_{\text{log}}(y) = \log y$ [Kuditipudi et al., 2024, Fernandez et al., 2023].



Simulation results for Gumbel-max watermark



Watermark detection 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 detection under text modification

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

Hypothesis testing under mixtures

$$H_0 : Y_t \sim \mu_0 \quad \forall t \quad \text{vs} \quad H_1^{\text{mix}} : Y_t | \mathbf{P}_t \sim (1 - \varepsilon_n) \mu_0 + \varepsilon_n \mu_{1, \mathbf{P}_t} \quad \forall t$$

- Difficulties: We know nothing about ε_n or $\mathbf{P}_t$.

Watermark detection under text modification

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

Hypothesis testing under mixtures

$$H_0 : Y_t \sim \mu_0 \quad \forall t \quad \text{vs} \quad H_1^{\text{mix}} : Y_t | \mathbf{P}_t \sim (1 - \varepsilon_n) \mu_0 + \varepsilon_n \mu_{1, \mathbf{P}_t} \quad \forall t$$

- Difficulties: We know nothing about ε_n or $\mathbf{P}_t$.

Hope: We know everything about the null H_0 which always assume $Y_{1:n}$ iid from μ_0 .

Watermark detection under text modification

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

Hypothesis testing under mixtures

$$H_0 : Y_t \sim \mu_0 \quad \forall t \quad \text{vs} \quad H_1^{\text{mix}} : Y_t | \mathbf{P}_t \sim (1 - \varepsilon_n)\mu_0 + \varepsilon_n\mu_{1, \mathbf{P}_t} \quad \forall t$$

- Difficulties: We know nothing about ε_n or $\mathbf{P}_t$.

Hope: We know everything about the null H_0 which always assume $Y_{1:n}$ iid from μ_0 .

- Focus to determine whether the observed $Y_1, \dots, Y_n$ follows μ_0 .
- Tr-GoF [Li et al., 2025a] checks the deviation between the empirical CDF of $Y_{1:n}$ and μ_0 via an f -divergence.
- Too large deviation indicate the existence of watermarked subtexts.

- The empirical CDF of p-values: $\mathbb{F}_n(r) = \frac{1}{n} \sum_{t=1}^n \mathbf{1}_{p_t \leq r}$ where $p_t = 1 - F_0(Y_t)$.
- Under H_0 , we must have $p_1, \dots, p_n$ i.i.d. from $\mathcal{U}(0, 1)$.
- Tr-GoF rejects H_0 if $\sup_{r \in (p_1, 1)} nK_s(\mathbb{F}_n(r), r)$ is larger than expected for $s \in [-1, 2]$

where

$$K_s(u, v) = \frac{1}{s(1-s)} \left[1 - u^s v^{1-s} - (1-u)^s (1-v)^{1-s} \right].$$

- One can show that K_s is the ϕ_s -divergence between two Bernoulli variables:

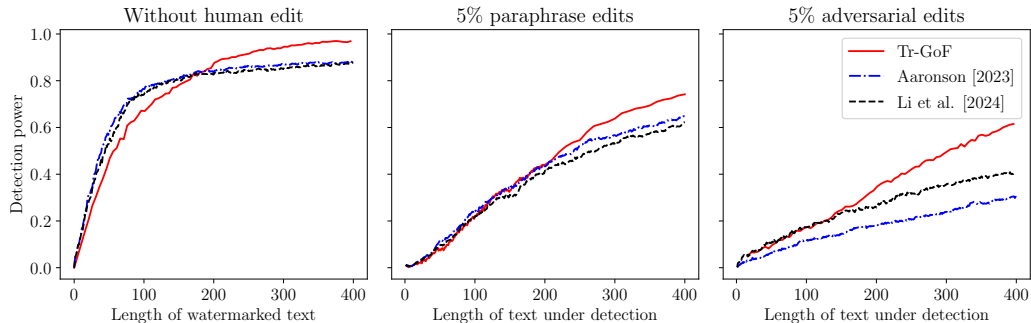
$$K_s(u, v) = D_{\phi_s}(\text{Ber}(u) \parallel \text{Ber}(v)) = v\phi_s\left(\frac{u}{v}\right) + (1-v)\phi_s\left(\frac{1-u}{1-v}\right),$$

where ϕ_s is a scalar convex function indexed by s :

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

Robust performance of Tr-GoF

- C4 news-like dataset [Raffel et al., 2020] & OPT-1.3B model [Zhang et al., 2022].



Why Tr-GoF performs so well? [Li et al., 2025a]

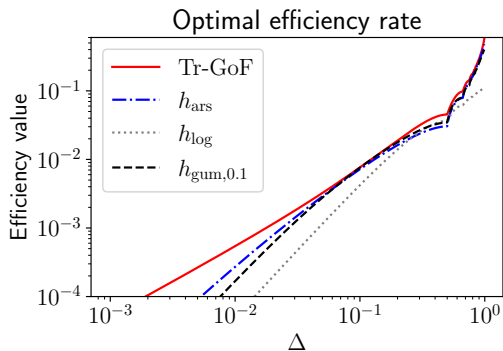
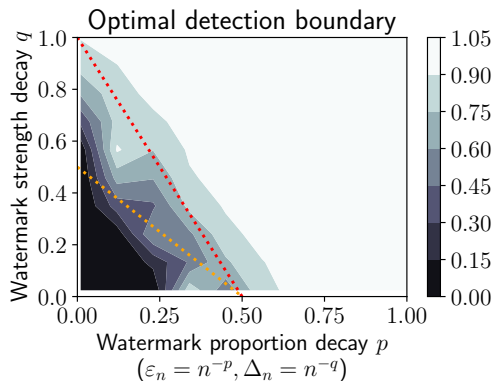
A theoretical question

Why Tr-GoF performs so well in the mixture watermark detection problem?

Why Tr-GoF performs so well? [Li et al., 2025a]

A theoretical question

Why Tr-GoF performs so well in the mixture watermark detection problem?

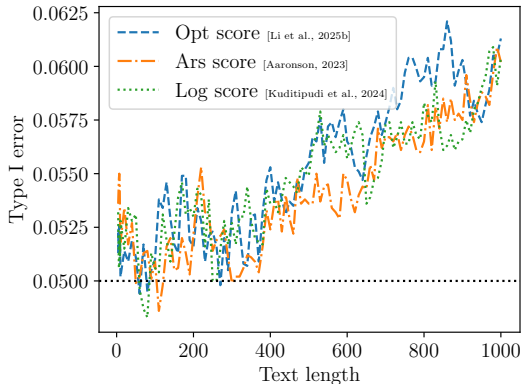
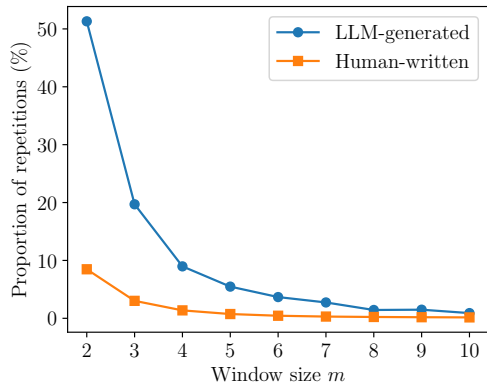


Optimal detection under pseudorandom collision [Cai et al., 2025]

- **Perfect pseudorandomness:** treats $\zeta_{1:n}$ as i.i.d., often unrealistic.
- Each $\zeta_t = \mathcal{A}(w_{(t-m):(t-1)}, \text{Key})$ uses an m -token text window.

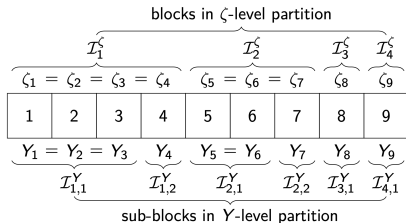
Optimal detection under pseudorandom collision [Cai et al., 2025]

- **Perfect pseudorandomness:** treats $\zeta_{1:n}$ as i.i.d., often unrealistic.
- Each $\zeta_t = \mathcal{A}(w_{(t-m):(t-1)}, \text{Key})$ uses an m -token text window.



Optimal detection under pseudorandom collision [Cai et al., 2025]

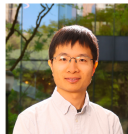
- **Perfect pseudorandomness:** treats $\zeta_{1:n}$ as i.i.d., often unrealistic.
- Each $\zeta_t = \mathcal{A}(w_{(t-m):(t-1)}, \text{Key})$ uses an m -token text window.
- We propose a two-layer hierarchy partition to capture collision patterns and derive the optimal detector within it.
- **Takeaway:** We theoretically justify the practice: for Gumbel-max, just remove repeated Y_t 's and apply $h_{\text{gum}, \Delta}$.



T. Tony Cai



Qi Long



Weijie Su



Garrett G. Wen

Summary

Watermarking

- A principled statistical framework for detecting unbiased watermarks.
- Optimal detection remains possible under user corruption and imperfect pseudorandomness.

Part II: Evaluation of Unseen LLM Capabilities

Evaluating the Unseen Capabilities: How Many Theorems Do LLMs Know?

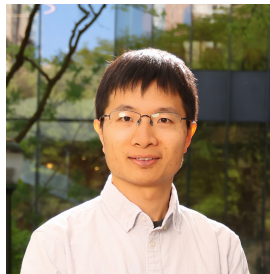
Under review at PNAS



Jiayi Xin

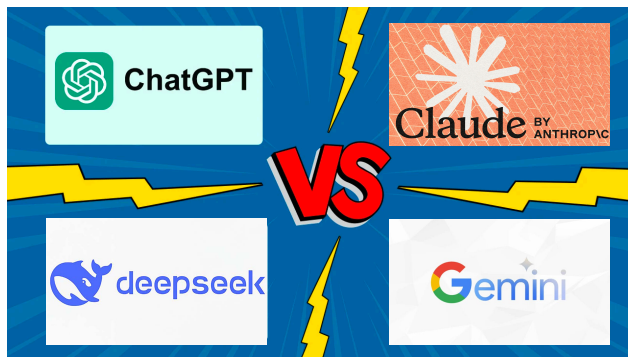


Qi Long



Weijie Su

Evaluation of LLMs



- Validate performance of reliability of LLMs.
- Guide effective and trustworthy deployment.
- Reveal potential limitations, biases, or vulnerabilities.

An evaluation crisis



Andrej Karpathy ✓

@karpathy

Following

Building @EurekaLabsAI. Previously Director of AI @ Tesla, founding team @ OpenAI, CS231n/PhD @ Stanford. I like to train large deep neural nets.

My reaction is that there is an evaluation crisis. I don't really know what metrics to look at right now.

MMLU was a good and useful for a few years but that's long over. SWE-Bench Verified (real, practical, verified problems) I really like and is great but itself too narrow.

Chatbot Arena received so much focus (partly my fault?) that LLM labs have started to really overfit to it, via a combination of prompt mining (from API requests), private evals bombardment, and, worse, explicit use of rankings as training supervision. I think it's still ~ok and there's a lack of "better", but it feels on decline in signal.

There's a number of private evals popping up, an ensemble of which might be one promising path forward.

In absence of great comprehensive evals I tried to turn to vibe checks instead, but I now fear they are misleading and there is too much opportunity for confirmation bias, too low sample size, etc., it's just not great.

TLDR my reaction is I don't really know how good these models are right now.

1:29 PM · Mar 2, 2025 · 301.9K Views

An evaluation crisis



Andrej Karpathy ✓

@karpathy

Following

Building @EurekaLabsAI. Previously Director of AI @ Tesla, founding team @ OpenAI, CS231n/PhD @ Stanford. I like to train large deep neural nets.

Some causes of the crisis

- Benchmark contamination [Sainz et al., 2023].
- Overfitting through repeated leaderboard submissions [Singh et al., 2025].
- Narrow test-time optimization strategies [Leech et al., 2024].

My reaction is that there is an evaluation crisis. I don't really know what metrics to look at right now.

MMLU was a good and useful for a few years but that's long over. SWE-Bench Verified (real, practical, verified problems) I really like and is great but itself too narrow.

Chatbot Arena received so much focus (partly my fault?) that LLM labs have started to really overfit to it, via a combination of prompt mining (from API requests), private evals bombardment, and, worse, explicit use of rankings as training supervision. I think it's still ~ok and there's a lack of "better", but it feels on decline in signal.

There's a number of private evals popping up, an ensemble of which might be one promising path forward.

In absence of great comprehensive evals I tried to turn to vibe checks instead, but I now fear they are misleading and there is too much opportunity for confirmation bias, too low sample size, etc., it's just not great.

TLDR my reaction is I don't really know how good these models are right now.

1:29 PM · Mar 2, 2025 · 301.9K Views

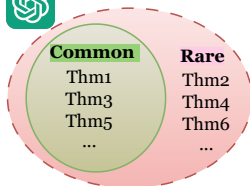
An alternative perspective for LLM evaluation

Could we evaluate LLMs by estimating their “unseen” capacity or knowledge?

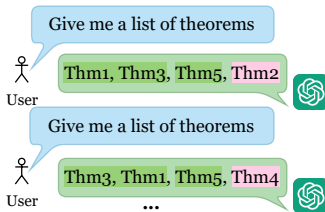
An alternative perspective for LLM evaluation

Could we evaluate LLMs by estimating their “unseen” capacity or knowledge?

(a) How Many Theorems Does an LLM Know?

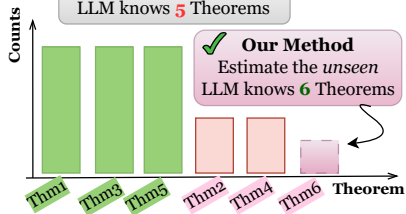


(b) Prompt 100 times

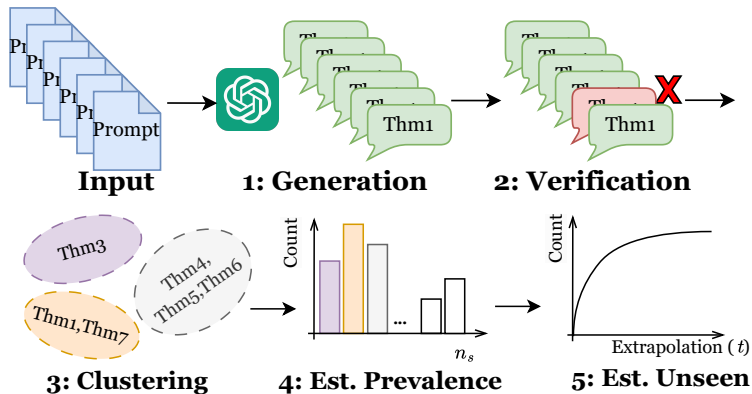


(c)

X From Observation
LLM knows 5 Theorems

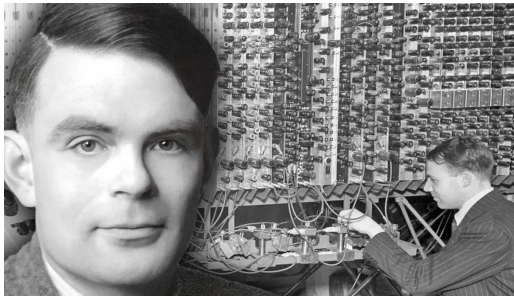


Our method: KnowSum

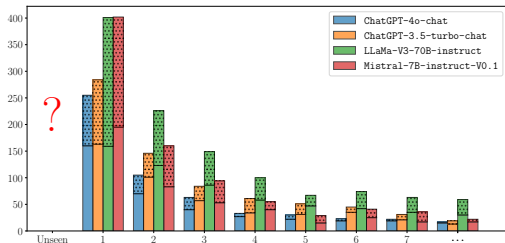


- Difficulties: hallucination & redundancy.
- Steps 2 & 3 standardize responses.
- Prevalence n_s = count of tems appearing s times ($s \geq 1$).

How to extrapolate from seen to unseen



Good-Turing (GT) estimator for cracking Enigma



Biometrika (1976), 63, 3, pp. 435–47

With 3 text-figures

Printed in Great Britain

435

Estimating the number of unseen species: How many words did Shakespeare know?

BY BRADLEY EFRON AND RONALD THISTED

Department of Statistics, Stanford University, California

SUMMARY

Shakespeare wrote 31 534 different words, of which 14 376 appear only once, 4343 twice, etc. The question considered is how many words he knew but did not use. A parametric empirical Bayes model due to Fisher and a nonparametric model due to Good & Toulmin are examined. The latter theory is augmented using linear programming methods. We conclude that the models are equivalent to supposing that Shakespeare knew at least 35 000 more words.

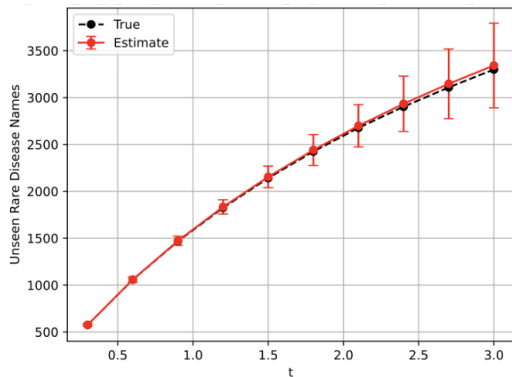
Smoothed GT estimator for estimating vocabulary

$$\hat{N}_{\text{unseen}}^{\text{ET}}(t) = \mathbb{E} \left[- \sum_{s=1}^L (-t)^s n_s \right]$$

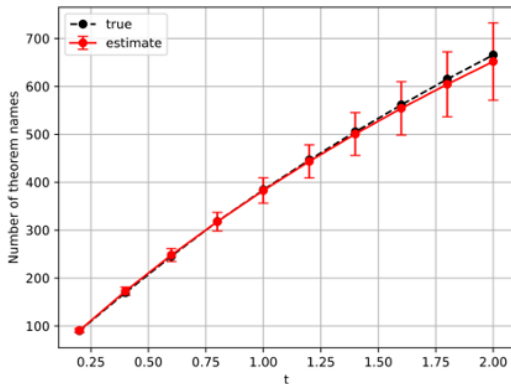
How to extrapolate from seen to unseen

Llama 3 70B Instruct

How many rare diseases



How many theorems



Unseen portion

How many theorems do LLMs know?

Seen



868



753



702



444

How many theorems do LLMs know?

Seen



868



753



702



444

Total



1194



1189



1162



1064

Summary

Watermarking

- A principled statistical framework for detecting unbiased watermarks.
- Optimal detection remains possible under user corruption and imperfect randomness.

Model Evaluation

- Estimating unseen knowledge enables more faithful LLM evaluation.
- Accounting for unseen knowledge can change model rankings.

What can statistics offer to LLMs?

What can statistics offer to LLMs?

From practice to theory: Practical needs motivate new statistical questions.

- Optimal detection and robust inference for watermarking signals.

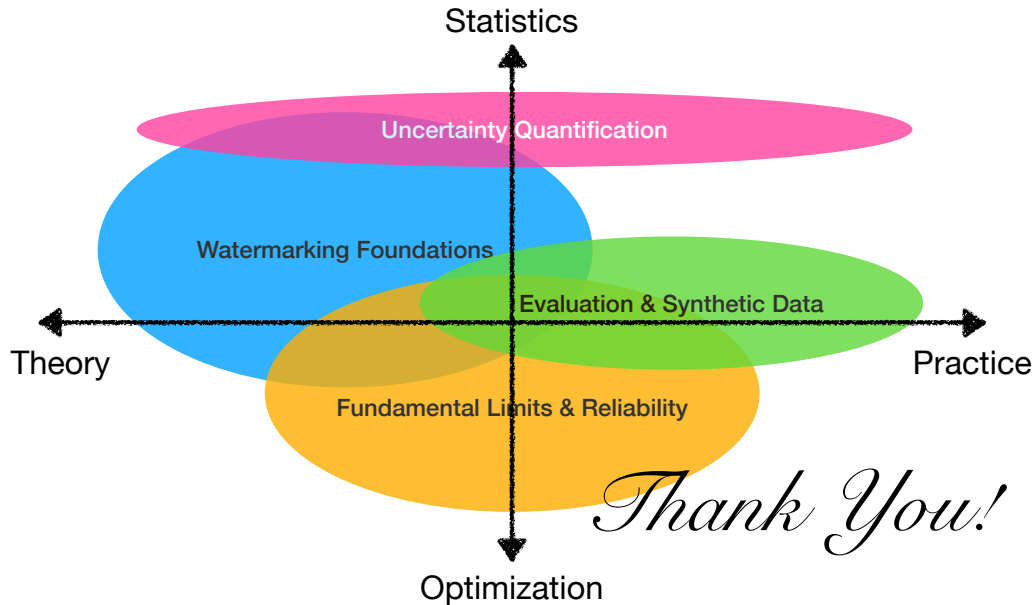
From theory to practice: Statistical ideas provide insights into LLM behavior.

- Good–Turing estimators reveal unseen LLM knowledge.

Another philosophy: Interplay between Stat and Opt

- Statistical machine learning
 - Non-stationary data [Li et al., 2023a]
 - Heavy-tailed noise [Li and Sun, 2024]
 - Adversarial corruption [Yu et al., 2025]
 - High-dimensionality [Li et al., 2025f]
 - Sparsity [Yang et al., 2019]
- Stochastic approximation
 - Polyak averaging [Li et al., 2023b]
 - Second-order method [Li et al., 2020b]
 - Two-time-scale [Han et al., 2024b]
 - Functional central limit theorem [Han et al., 2024a]
 - Jump diffusion [Liang et al., 2022]
- Federated optimization
 - Convergence analysis [Li et al., 2020a]
 - Confidence intervals [Li et al., 2022]
 - Communication efficiency [Li et al., 2019, 2021]
 - Personalization [Lin et al., 2022, Han et al., 2024c]
 - Privacy [Guo et al., 2024, Li et al., 2025c]

Future directions



Backup Slides

Joint null conditional distribution of (w_t, ζ_t)

(A1) Under H_0 , $w_t \perp \zeta_t$ conditional on $(w_{1:(t-1)}, \zeta_{1:(t-1)})$.

(A2) ζ_t is (statistically) independent of $w_{1:(t-1)}$

(A3) $\zeta_{1:n}$ are i.i.d. copies of a random variable with law $\mathbb{P}_\xi$.

$$\begin{aligned}\mathbb{P}_0(w_t, \zeta_t \mid w_{1:(t-1)}, \zeta_{1:(t-1)}) &\stackrel{(A1)}{=} \mathbb{P}_0(w_t \mid w_{1:(t-1)}, \zeta_{1:(t-1)}) \cdot \mathbb{P}(\zeta_t \mid w_{1:(t-1)}, \zeta_{1:(t-1)}) \\ &\stackrel{\zeta_t = \mathcal{A}(w_{1:(t-1)}, \text{Key})}{=} \mathbb{P}_0(w_t \mid w_{1:(t-1)}) \cdot \mathbb{P}(\zeta_t \mid w_{1:(t-1)}) \\ &\stackrel{(A2)+(A3)}{=} P_t \cdot \mathbb{P}_\zeta(\zeta_t).\end{aligned}$$

Why pivotal statistics are i.i.d. under H_0

1. Under H_0 ,

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

2. By definition of a pivotal statistic,

$$Y_t := Y(w_t, \zeta_t) \text{ satisfies } Y_t \mid (w_{1:(t-1)}, \zeta_{1:(t-1)}) \sim \mu_0.$$

3. Since $\sigma(Y_{1:(t-1)}) \subset \sigma(w_{1:(t-1)}, \zeta_{1:(t-1)})$, the tower property gives, for any bounded measurable f ,

$$\mathbf{E}[f(Y_t) \mid Y_{1:(t-1)}] = \mathbf{E}\left[\mathbf{E}[f(Y_t) \mid w_{1:(t-1)}, \zeta_{1:(t-1)}] \mid Y_{1:(t-1)}\right] = \int f d\mu_0.$$

4. Hence $Y_t \mid Y_{1:(t-1)} \sim \mu_0$, and $\{Y_t\}_{t \geq 1}$ are i.i.d. under H_0 .

Hypothesis testing under mixtures

$$H_0 : Y_t \sim \mu_0 \quad \forall t \quad \text{versus} \quad H_1^{\text{mix}} : Y_t | \mathbf{P}_t \sim (1 - \varepsilon_n) \mu_0 + \varepsilon_n \mu_{1, \mathbf{P}_t} \quad \forall t.$$

Definition (*A difficult case*)

We consider an extreme case where

- $\varepsilon_n \asymp n^{-p}$ for all t and $p \in (0, 1]$.
- $1 - \max_{w \in \mathcal{W}} \mathbf{P}_{t,w} = \Delta_n$ for all t with $\Delta_n \asymp n^{-q}$ and $q \in (0, 1)$.
- Motivated by sparse detection problem [Donoho and Jin, 2004, 2015].
- If $\varepsilon_n = 0$ or $1 - \max_{w \in \mathcal{W}} \mathbf{P}_{t,w} = 0$, $(1 - \varepsilon_n) \mu_0 + \varepsilon_n \mu_{1, \mathbf{P}_t} = \mu_0$, i.e., H_0 merges with H_1^{mix} .

When the robust detection is possible?

Optimal detection boundary

Theorem (*Optimal detection boundary on the (p, q) -plane*)

- If $q + 2p > 1$, H_0 and H_1^m merge asymptotically. For any test, the sum of Type I and Type II error probabilities is 1 as $n \rightarrow \infty$.
- If $q + 2p < 1$, H_0 and H_1^m separate asymptotically. Furthermore, for the likelihood-ratio test that rejects H_0 if the log-likelihood ratio is positive, the sum of Type I and Type II error probabilities tends to 0 as $n \rightarrow \infty$.

$\Rightarrow$ Robust detection is impossible for small watermark signal, i.e., $q + 2p > 1$.

$\Rightarrow$ With sufficient watermark signal, detection is possible with the likelihood-ratio test an optimal rule, i.e., $q + 2p < 1$.

!!! The likelihood-ratio test is impractical as it needs to know $\mathbf{P}_t$'s and ε_n .

Target

An ideal optimal detection method should work as long as $q + 2p < 1$ and don't requires the knowledge of $\mathbf{P}_t$'s and ε_n .

Our finding

The GoF test achieves this optimal detection boundary.

Theorem (*Adaptive optimality*)

The Type I and II errors of the Tr-GoF test $\rightarrow 0$ if $n \rightarrow \infty$ as long as $q + 2p < 1$.

- Optimal adaptivity without any prior knowledge.

What about constant corruption?

Optimal detection efficiency

- The optimal detection boundary cares about the diminishing region where the watermark signal decays with the text length n .
- Practical settings meet with the constant corruption case, i.e., $\varepsilon_n \equiv \varepsilon$.
- The problem is detectable because $p = q = 0$ (within $q + 2p < 1$).
- Recall $\mathcal{P}$ -efficiency: the rate of exponential decrease in Type II errors for a fixed significance level α and the worst-case alternative within a belief set $\mathcal{P}$.

Definition ($\mathcal{P}$ -efficiency [Li et al., 2025b], revisited)

Let $\gamma_{n,\alpha}$ satisfy $\mathbb{P}_0(S_n \geq \gamma_{n,\alpha}) = \alpha$ for $n \geq 1$. For a given belief set $\mathcal{P}$, we define the following limit (if exists) as the $\mathcal{P}$ -efficiency of S_n and denote it by $R_{\mathcal{P}}(S_n)$:

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{P}_t \in \mathcal{P}, \forall t \in [n]} \frac{1}{n} \log \mathbb{P}_1(S_n \leq \gamma_{n,\alpha}) = -R_{\mathcal{P}}(S_n).$$

What about constant corruption?

Optimal detection efficiency

Theorem (*Optimal $\mathcal{P}_\Delta$ -efficiency*)

Let $s \in (0, 1)$, $\varepsilon_n \equiv \varepsilon \in (0, 1]$ and $\Delta_n \equiv \Delta \in (0, 1)$.

$$R_{\mathcal{P}_\Delta}(\text{any detection rule}) \leq D_{\text{KL}}(\mu_0, (1 - \varepsilon)\mu_0 + \varepsilon\mu_{1, \mathbf{P}_\Delta^*}) \leq R_{\mathcal{P}_\Delta}(\text{Tr} - \text{GoF})$$

where $\mathbf{P}_\Delta^*$ is the least-favorable NTP distribution defined by

$$\mathbf{P}_\Delta^* = \left(\underbrace{1 - \Delta, \dots, 1 - \Delta}_{\lfloor \frac{1}{1-\Delta} \rfloor \text{ times}}, 1 - (1 - \Delta) \cdot \left\lfloor \frac{1}{1 - \Delta} \right\rfloor, 0, \dots \right).$$

- Upper and lower bounds.
- When $\varepsilon = 1$, this rate is obtained by the sum-based test defined by $h_{\text{gum}, \Delta}$.
- Optimal efficiency without any prior knowledge.

References I

- S. Aaronson. Watermarking of large language models.
<https://simons.berkeley.edu/talks/scott-aaronson-ut-austin-openai-2023-08-17>, August 2023.
- P. Abdalla and R. Vershynin. LLM watermarking using mixtures and statistical-to-computational gaps. *arXiv preprint arXiv:2505.01484*, 2025.
- 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.
- B. Barak. An intensive introduction to cryptography, lectures notes for Harvard CS 127.
<https://intensecrypto.org/public/index.html>, Fall 2021.
- T. T. Cai, X. Li, Q. Long, W. J. Su, and G. G. Wen. Optimal detection for language watermarks with pseudorandom collision. *arXiv preprint arXiv:2510.22007*, 2025.
- Y. Chen, Z. Hu, Y. Wu, R. Chen, Y. Jin, M. Zhan, C. Xie, W. Chen, and H. Huang. Enhancing privacy in biosecurity with watermarked protein design. *Bioinformatics*, page btaf141, 2025.
- M. Christ, S. Gunn, and O. Zamir. Undetectable watermarks for language models. In *Conference on Learning Theory*, pages 1125–1139. PMLR, 2024.
- S. Dathathri, A. See, S. Ghaisas, P.-S. Huang, R. McAdam, J. Welbl, V. Bachani, A. Kaskasoli, R. Stanforth, T. Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634 (8035):818–823, 2024.
- D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

References II

- D. Donoho and J. Jin. Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical science*, 30(1):1–25, 2015.
- B. Emi and M. Spero. Technical report on the pangram AI-generated text classifier. *arXiv preprint arXiv:2402.14873*, 2024.
- 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.
- J. Fu, X. Zhao, R. Yang, Y. Zhang, J. Chen, and Y. Xiao. GumbelSoft: Diversified language model watermarking via the GumbelMax-trick. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5791–5808, 2024.
- T. Gloaguen, N. Jovanović, R. Staab, and M. Vechev. Towards watermarking of open-source LLMs. *arXiv preprint arXiv:2502.10525*, 2025.
- GPTZero. GPTZero: More than an AI detector preserve what's human. <https://gptzero.me/>, 2023.
- E. J. Gumbel. *Statistical theory of extreme values and some practical applications: A series of lectures*, volume 33. US Government Printing Office, 1948.
- X. Guo, X. Li, X. Chang, S. Wang, and Z. Zhang. Fedpower: Privacy-preserving distributed eigenspace estimation. *Machine Learning*, 113:8427–8458, 2024. Extended version of the conference paper: Communication-efficient distributed SVD via local power iterations.
- Y. Han, X. Li, J. Liang, and Z. Zhang. Decoupled functional central limit theorems for two-time-scale stochastic approximation. *arXiv preprint arXiv:2412.17070*, 2024a.

References III

- Y. Han, X. Li, and Z. Zhang. Finite-time decoupled convergence in nonlinear two-time-scale stochastic approximation. *arXiv preprint arXiv:2401.03893*, 2024b.
- Y. Han, S. Lin, X. Li, and Z. Zhang. A random projection approach to personalized federated learning: Enhancing communication efficiency, robustness, and fairness. *Journal of Machine Learning Research*, 25 (380):1–88, 2024c. Extended version of the conference paper: Personalized federated learning towards communication efficiency, robustness and fairness.
- W. He, X. Li, T. Shang, L. Shen, W. J. Su, and Q. Long. On the empirical power of goodness-of-fit tests in watermark detection. In *Advances in neural information processing systems*, 2025a.
- W. He, X. L. Li, L. Shen, W. J. Su, and Q. Long. Break the trade-off between watermark strength and speculative sampling efficiency for language models. In *Submitted to The Fourteenth International Conference on Learning Representations*, 2025b. URL <https://openreview.net/forum?id=HA8vzzT6Ax>. under review.
- Z. Hu, L. Chen, X. Wu, Y. Wu, H. Zhang, and H. Huang. Unbiased watermark for large language models. In *International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=uWVC5FVidc>.
- J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- 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, 2023.

References IV

- 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. In *International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=DEJIDCmW0z>.
- 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. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=FpaCL1M02C>.
- G. Leech, J. J. Vazquez, N. Kupper, M. Yagudin, and L. Aitchison. Questionable practices in machine learning. *arXiv preprint arXiv:2407.12220*, 2024.
- X. Li and Q. Sun. Variance-aware decision making with linear function approximation with heavy-tailed rewards. *Transactions on Machine Learning Research*, 2024.
- X. Li, W. Yang, S. Wang, and Z. Zhang. Communication efficient decentralized training with multiple local updates. *Technical report, arXiv preprint arXiv:1910.09126*, 2019.
- X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020a.
- X. Li, S. Wang, and Z. Zhang. Do subsampled newton methods work for high-dimensional data? In *AAAI Conference on Artificial Intelligence*, volume 34, pages 4723–4730, 2020b.

References V

- X. Li, S. Wang, K. Chen, and Z. Zhang. Communication-efficient distributed svd via local power iterations. In *International Conference on Machine Learning*, 2021.
- X. Li, J. Liang, X. Chang, and Z. Zhang. Statistical estimation and online inference via local sgd. In *Conference on Learning Theory*, volume 178, pages 1613–1661. PMLR, 2022.
- X. Li, J. Liang, X. Chen, and Z. Zhang. Convergence and inference of Stream sgd, with applications to queueing systems and inventory control. *arXiv preprint arXiv:2309.09545*, 2023a.
- X. Li, W. Yang, J. Liang, Z. Zhang, and M. I. Jordan. A statistical analysis of polyak-ruppert averaged q-learning. In *International Conference on Artificial Intelligence and Statistics*, volume 206, 2023b.
- X. Li, F. Ruan, H. Wang, Q. Long, and W. J. Su. Robust detection of watermarks for large language models under human edits. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf056, 2025a.
- X. Li, F. Ruan, H. Wang, Q. Long, and W. J. Su. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *The Annals of Statistics*, 53(1):322–351, 2025b.
- X. Li, C. Wang, B. Su, Q. Long, and W. J. Su. Mitigating the privacy–utility trade-off in decentralized federated learning via f-differential privacy. In *Advances in Neural Information Processing Systems*, volume 38, 2025c.
- X. Li, G. Wen, W. He, J. Wu, Q. Long, and W. J. Su. Optimal estimation of watermark proportions in hybrid AI–human texts. *arXiv preprint arXiv:2506.22343*, 2025d.
- X. Li, J. Xin, Q. Long, and W. J. Su. Evaluating the unseen capabilities: How many theorems do llms know? *arXiv preprint arXiv:2506.02058*, 2025e.

References VI

- X. Li, Y. Zhang, J. Wu, Q. Long, and W. J. Su. High-dimensional nonparametric density estimation via max-random forest. *Manuscript available upon request*, 2025f.
- J. Liang, Y. Han, X. Li, and Z. Zhang. Asymptotic behaviors of projected stochastic approximation: A jump diffusion perspective. In *Advances in Neural Information Processing Systems*, 2022.
- W. Liang, M. Yuksekogonul, 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. In *International Conference on Machine Learning*, 2024.
- S. Lin, Y. Han, X. Li, and Z. Zhang. Personalized federated learning towards communication efficiency, robustness and fairness. In *Neural Information Processing Systems*, 2022.
- E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning*, pages 24950–24962. PMLR, 2023.
- L. Pan, A. Liu, Z. He, Z. Gao, X. Zhao, Y. Lu, B. Zhou, S. Liu, X. Hu, L. Wen, et al. MarkLLM: An Open-Source Toolkit for LLM Watermarking. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 61–71, 2024.
- J. Piet, C. Sitawarin, V. Fang, N. Mu, and D. Wagner. Mark my words: Analyzing and evaluating language model watermarks. *arXiv preprint arXiv:2312.00273*, 2023.

References VII

- C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- 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.
- O. Sainz, J. Campos, I. García-Ferrero, J. Etxaniz, O. L. de Lacalle, and E. Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, 2023.
- I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.
- S. Singh, Y. Nan, A. Wang, D. D’Souza, S. Kapoor, A. Üstün, S. Koyejo, Y. Deng, S. Longpre, N. A. Smith, B. Ermiş, M. Fadaee, and S. Hooker. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*, 2025.
- W. Su. Do large language models (really) need statistical foundations? *arXiv preprint arXiv:2505.19145*, 2025.
- D. Weber-Wulff, A. Anohina-Naumeca, S. Bjelobaba, T. Foltýnek, J. Guerrero-Dib, O. Popoola, P. vSigt, and L. Waddington. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1):26, 2023.
- Y. Wu, Z. Hu, J. Guo, H. Zhang, and H. Huang. A resilient and accessible distribution-preserving watermark for large language models. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=c8qWiNiqRY>.

References VIII

- Y. Xie, X. Chen, Z. Ren, and W. J. Su. Watermark in the classroom: A conformal framework for adaptive ai usage detection. *arXiv preprint arXiv:2507.23113*, 2025a.
- Y. Xie, X. Li, T. Mallick, W. Su, and R. Zhang. Debiasing watermarks for large language models via maximal coupling. *Journal of the American Statistical Association*, (just-accepted):1–21, 2025b.
- J. Xin, E. Qiang, X. Li, W. J. Su, and Q. Long. Paraphrase-robust conformal prediction for reliable LLM uncertainty quantification. In *Submitted to The Fourteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=Uf04r8gDn7>. under review.
- W. Yang, X. Li, and Z. Zhang. A regularized approach to sparse optimal policy in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- 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.
- Q. Yu, E. Baek, X. Li, and Q. Sun. Corruption-robust variance-aware algorithms for generalized linear bandits under heavy-tailed rewards. In *Conference on Uncertainty in Artificial Intelligence*, 2025.
- ZeroGPT. ZeroGPT: Trusted GPT-4, ChatGPT and AI detector tool by ZeroGPT. <https://www.zerogpt.com/>, 2023.
- S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

References IX

- X. Zhao, P. V. Ananth, L. Li, and Y.-X. Wang. Provable robust watermarking for AI-generated text. In *International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=SsmT8a045L>.
- X. Zhao, L. Li, and Y.-X. Wang. Permute-and-Flip: An optimally stable and watermarkable decoder for LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL <https://openreview.net/forum?id=YyVVicZ32M>.
- Y. Zhao, X. Li, P. Song, Q. Long, and W. J. Su. TAB-DRW: A DFT-based robust watermark for generative tabular data. In *Submitted to The Fourteenth International Conference on Learning Representations*, 2025b. URL <https://openreview.net/forum?id=cmmu25plng>. under review.