On the Empirical Power of Goodness-of-Fit Tests in Watermark Detection

Weiqing He* Xiang Li* Tianqi Shang Li Shen Weijie Su Qi Long

University of Pennsylvania

Why use GoF tests for watermark detection

Watermark generation and detection. In watermarked text generation, each token $w_t \sim \boldsymbol{P}_t$ is drawn by a decoding function $\mathcal{S}$ as $w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t)$, where $\boldsymbol{P}_t$ is the next-token predictive distribution. This process induces a watermark signal—**a statistical dependence between w_t and ζ_t**—absent in human-written text, since humans lack access to ζ_t. Detection methods exploit this difference via pivotal statistics [1], defined as $Y_t = Y(w_t, \zeta_t)$, within a hypothesis-testing framework:

$$H_0 : w_{1:n} \text{ is human-written} \quad \text{vs.} \quad H_1 : w_{1:n} \text{ is LLM-generated.} \tag{1}$$

Property of Y: $Y(w, \zeta) \sim \mu_0$ whenever w and ζ are independent—as in human-written text—yielding the equivalent problem:

$$H_0 : Y_t \sim \mu_0 \text{ i.i.d.} \quad \text{vs.} \quad H_1 : Y_t \sim \text{a distribution dependent on } \boldsymbol{P}_t. \tag{2}$$

🤔 **Motivation**: Goodness-of-fit (GoF) tests are classical tools in statistics for determining whether a sequence of i.i.d. samples comes from a specified distribution, and the formulation is highly relevant to watermark detection!

- Rich literature on GoF but almost no application on watermark.
- Reveal the limitations of existing detection methods.
- Inspire the development of new detection methods.

🔍 Key Insights

GoF tests are a simple yet powerful tool for watermark detection!

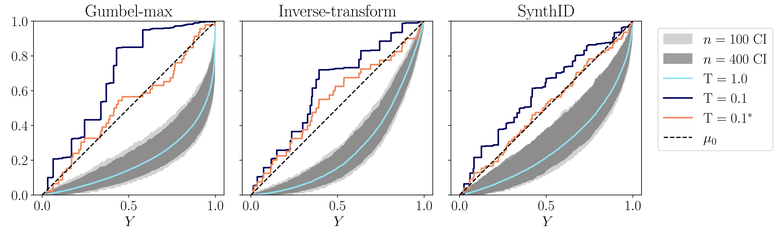
- GoF tests are highly effective at detecting watermark signals at high temperatures.
- Text repetition increases the deviation between the empirical CDF and the null distribution μ_0, enabling GoF tests to outperform baseline methods in low-temperature settings.
- GoF tests are robust to common edits and information-rich edits.

Detection Procedure

Require: Token sequence $w_{1:n}$; watermark decoder $\mathcal{S}$; significance level α; CDF under the null F_0.

1: **Compute Y_t and p-values**: $\mathrm{p}_t = 1 - F_0(Y_t),\ t = 1, \ldots, n$ from $w_{1:n}$.
2: **Sort** $\mathrm{p}_{(1)} \leq \cdots \leq \mathrm{p}_{(n)}$.
3: **Compute test statistic:** $D_n \leftarrow \max_{1 \leq i \leq n} \max\left(\mathrm{p}_{(i)} - \frac{i-1}{n}, \frac{i}{n} - \mathrm{p}_{(i)}\right)$.
4: **if** $D_n > \gamma_\alpha$ **then**: **Reject** H_0.
5: **else**: **Do not reject** H_0.
6: **end if**

Illustration of GoF tests



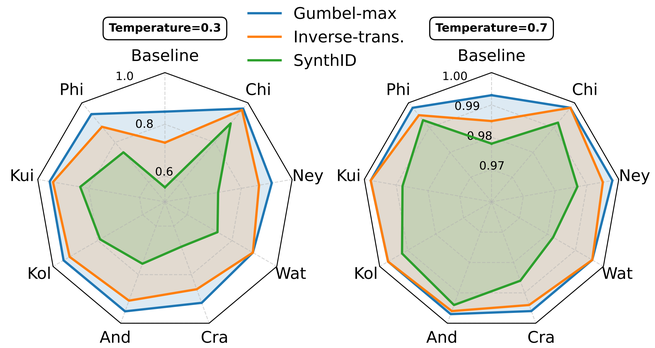
- GoF tests capture **distribution-level deviations**, typically by measuring the gap between the empirical and theoretical CDFs under the null hypothesis.
- GoF tests are particularly effective at high temperatures. The CDF difference is already apparent at $n = 100$, and increasing n further primarily reduces statistical variance.
- Repetition introduces step-like shapes of the CDF at low temperature that pushes the distribution away from the null.

Experiment Settings

- 3 popular watermarks: Gumbel-max, Inverse Transform, and SynthID.
- 3 open-source LLMs: OPT-1.3B, OPT-13B, and Llama 3.1-8B.
- 4 temperature settings: $\mathrm{T} \in \{0.1, 0.3, 0.7, 1.0\}$.
- 2 text generation tasks: text completion (`C4` dataset) and long-form question answering (`ELI5` dataset).

Detection Performance

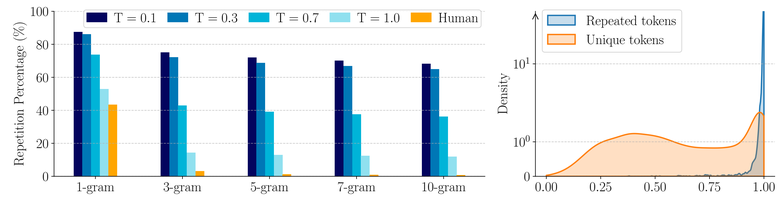
TPR @ 1% FPR



Understand the power of GoF tests

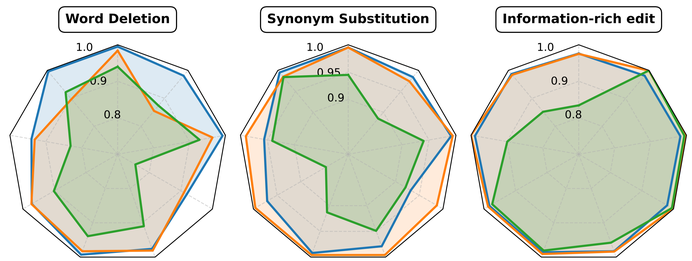
🔥 **At high temperatures**: GoF can utilize the full empirical distribution (CDF) of the pivotal statistics to detect deviations from the null μ_0. In contrast, baseline methods typically rely on sum-based statistics, which compress the data into a single value (reject H_0 if the sum $\sum_{t=1}^{n} h(Y_t)$ exceeds a critical threshold).

❄️ **At low temperatures**: LLM tends to generate more repeated text, which reinforces low-entropy behavior and leads to repeated pivotal statistics Y_t. Repeated Y_t pushes the empirical CDF away from the null.



💡 **GoF tests are effective at capturing CDF differences!**

Robustness Evaluation



😈 **Information-rich edits.** We consider a stronger editing setting in which the hash function $\mathcal{A}$ and secret key `Key` are known to the user. In this case, the user can selectively modify a limited number of LLM-generated tokens to reduce watermark signals while preserving the overall quality of the text.

👈 The legend and axis labels are identical to those in the left plot.

References

[1] Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie 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, 2025.