NEURAL INFORMATION PROCESSING SYSTEMS

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

Speaker: Weiqing He

Penn
University of Pennsylvania

NEURAL INFORMATION PROCESSING SYSTEMS

Background

- What's the main problem? - **Distinguish Human-written and LLM-generated content**

- What is watermark? - Currently, the main approach to solve this problem, which **secretly** modifies the output of LLM and detectors then try to **recognize the modification**.

- How to detect watermarks?
  - **Hypothesis test problem**:
    - $H_0 : Y_t \overset{i.i.d.}{\sim} \mu_0 \forall t \in [n]$ versus $H_1 : Y_t$ follows other distribution which depends on P_t

Human-written *LLM-generated*

Penn
University of Pennsylvania

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- The fundamental idea of Goodness of fit (GoF) tests is very relevant to the watermark detection problem and these tests can be adapted to the detection!

$$H_0 : Y_t \sim \mu_0 \text{ i.i.d. } \forall t \text{ vs. } H_1 : Y_t \sim \mu_1 \text{i.i.d.} \forall t$$

- Rich literature on GoF but almost no application on watermark. (Sum-based methods from previous studies, i.e. $\sum_{t=1}^{n} h(Y_t)$)
- Reveal the limitations of existing detection methods and inspire the development of new detection method.

Penn
University of Pennsylvania

NEURAL INFORMATION PROCESSING SYSTEMS

Results and findings

Experiment settings

- 3 watermarks: Gumbel-max, Inverse-transform, SynthID
- 8 GoF tests.
- 3 Open-source LLMs: OPT1.3B, OPT-13B, Llama 3.1-8B
- 4 temperatures: 0.1, 0.3, 0.7, 1.0
- 2 text generation tasks: Text completion from C4 and Long-form question answering from ELI5
- 4 types of edits: word deletion, synonym substitution, information-rich edit
- Detection baselines: Sum-based methods from previous studies, i.e. $\sum_{t=1}^{n} h(Y_t)$

Penn
University of Pennsylvania

NEURAL INFORMATION PROCESSING SYSTEMS

Results and findings

Type I error control and Type II error decay

- In watermark detection, a Type I error occurs when a non-watermarked text is falsely identified as watermarked, while a Type II error occurs when a genuine watermark is incorrectly rejected.

For GoF tests:

- Type I error is well controlled ($\alpha = 0.01$)
- Advantages are more pronounced at higher temperature settings.
- Advantages persist under low-temperature settings.

Table 2: Type I errors on human data and Type II errors (averaged over three LLMs) on the C4 dataset. All values are enlarged by 100 for readability. T denotes temperature and n the token length. Baseline refers to the best-performing method among the baseline detectors. Red shading highlights lower values; blue indicates higher values.

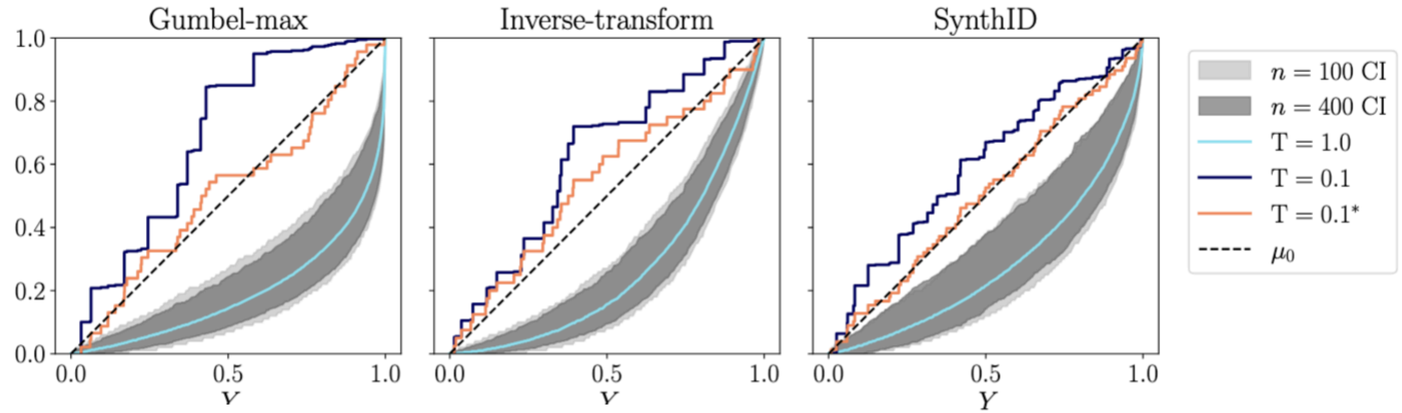
	T	n	Baseline	Phi	Kui	Kol	And	Cra	Wat	Ney	Chi
Gumbel-max	0.3	200	18.5	21.0	26.3	19.5	**15.5**	21.2	36.8	19.7	18.5
		400	15.1	5.7	4.7	4.7	4.9	8.4	10.7	8.0	**2.9**
	0.7	200	0.6	**0.3**	0.5	0.6	0.5	0.7	0.9	0.5	**0.3**
		400	0.7	**0.2**	**0.2**	0.3	**0.2**	0.4	0.4	**0.2**	**0.2**
	Type I		-	0.4	0.9	1.5	0.6	0.7	1.2	1.1	0.9
Inverse-tran.	0.3	200	38.7	51.0	29.2	29.7	33.6	36.7	40.4	37.3	**21.8**
		400	27.1	12.1	6.0	7.4	9.3	14.0	10.7	13.0	**3.6**
	0.7	200	1.3	2.7	1.3	**0.9**	**0.9**	1.0	1.9	1.2	2.3
		400	1.5	0.5	**0.2**	0.3	0.4	0.6	0.4	0.5	**0.2**
	Type I		-	0.4	1.4	1.2	1.0	1.0	1.4	1.5	0.6
SynthID	0.3	200	58.8	61.6	49.8	53.0	53.5	57.2	61.3	57.4	**36.4**
		400	44.4	25.0	16.7	21.0	24.5	31.5	26.5	29.1	**10.4**
	0.7	200	2.8	3.4	4.5	2.8	**2.3**	3.3	8.4	3.0	3.3
		400	2.2	0.7	1.2	0.8	**0.6**	1.4	1.8	1.3	0.8
	Type I		-	0.9	1.2	0.9	1.1	1.0	1.4	1.0	1.2

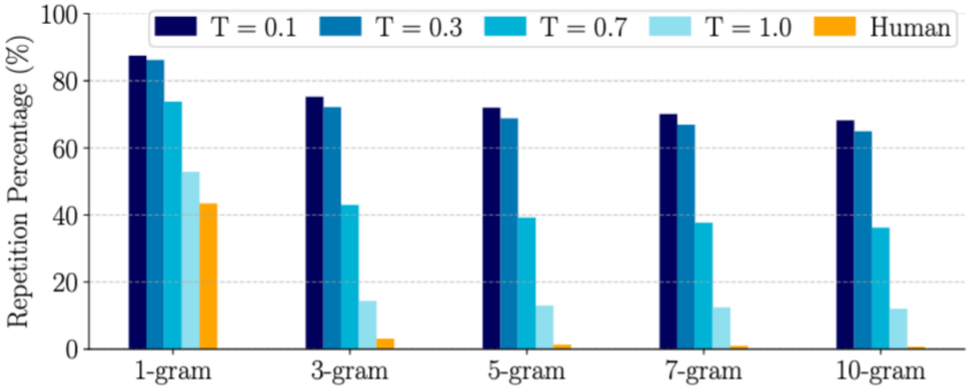
NEURAL INFORMATION PROCESSING SYSTEMS

Results and findings

Why do GoF tests perform well?

- For high temperatures:
  - baseline methods typically rely on sum-based statistics
  - GoF Use more “Features” for comparison.
- For low temperatures:
  - LLM tends to generate more repetitions
  - Repetition shifts the empirical CDF
- **GoF tests are effective at capturing CDF differences**





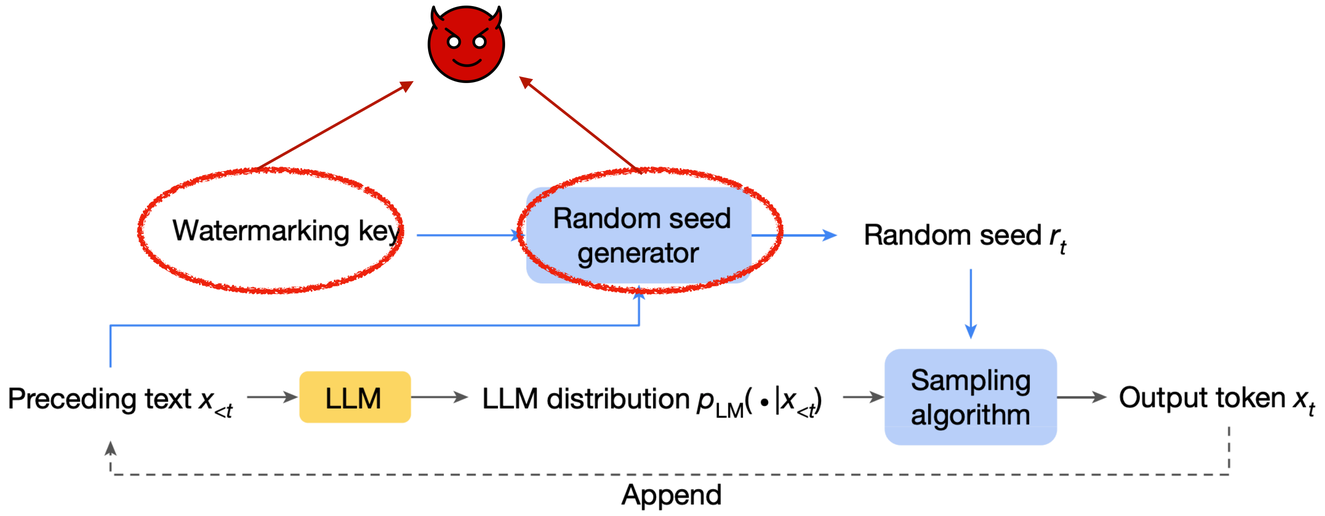
Penn
University of Pennsylvania

NEURAL INFORMATION PROCESSING SYSTEMS

Results and findings

Results on robustness

- Temperature is fixed at 1.0.
- 3 types of edits: word deletion, synonym substitution, information-rich edit
  - Information-rich edit: attacker knows random seed generator and watermarking key; replace the token with high pivotal statistics to avoid detection.



Penn
University of Pennsylvania

NEURAL INFORMATION PROCESSING SYSTEMS

Results and findings

Results on robustness

- For words deletion, synonym substitution, GoF tests' performances varies, but the best of them can be comparable to the baseline methods.
- For information-rich edit, GoF tests perform well.
  - Their strength lies in detecting deviations between the empirical CDF and the null u_0 rather than relying on extreme values of pivotal statistics.

Table 3: Type II errors (averaged over three LLMs) on the C4 dataset with temperature 1.0 under various editing types. "Del" denotes word deletion, "Sub" represents synonym substitution, and "Info" refers to information-rich edits. All values are enlarged by 100 for readability. `Baseline` refers to the best-performing method among the baseline detectors. Red shading highlights lower values; blue indicates higher values.

	Edits		Baseline	Phi	Kui	Kol	And	Cra	Wat	Ney	Chi
Gumbel-max	Del	@ 0.1	0.4	0.3	0.5	0.3	0.3	0.5	0.6	0.4	**0.2**
		@ 0.2	0.5	**0.4**	6.0	2.8	0.7	2.4	8.1	0.8	1.9
	Sub	@ 0.1	0.2	**0.1**	0.4	0.3	0.2	0.3	0.5	0.3	0.3
		@ 0.2	**0.5**	**0.5**	4.4	2.9	0.8	2.1	6.8	0.9	1.6
	Info	@ 0.3	2.4	1.3	**1.1**	1.7	1.4	1.7	2.0	1.7	1.9
		@ 0.5	38.0	34.7	**14.6**	33.3	29.7	30.8	24.2	28.2	26.8
Inverse-transform	Del	@ 0.1	0.3	0.9	0.6	0.2	**0.1**	**0.1**	1.0	0.3	1.6
		@ 0.2	**1.6**	10.3	6.9	2.7	1.7	1.9	9.0	3.6	14.5
	Sub	@ 0.1	0.2	0.3	**0.1**	**0.1**	**0.1**	**0.1**	0.2	**0.1**	0.2
		@ 0.2	0.5	1.5	1.0	**0.4**	**0.4**	**0.4**	1.3	0.6	2.6
	Info	@ 0.3	2.5	1.6	**0.0**	1.3	0.9	1.6	0.1	0.2	0.1
		@ 0.5	34.6	33.7	**0.4**	19.6	16.4	26.1	0.7	2.2	3.1
SynthID	Del	@ 0.1	2.9	**2.0**	3.2	2.4	**2.0**	2.5	4.7	2.1	2.1
		@ 0.2	**6.0**	7.8	16.9	9.8	6.1	9.0	24.4	7.2	12.6
	Sub	@ 0.1	2.3	1.6	2.7	1.8	**1.5**	1.9	4.1	1.6	1.8
		@ 0.2	5.5	5.9	15.3	8.7	**5.1**	7.9	21.9	6.0	11.5
	Info	@ 0.3	16.6	14.8	0.1	2.6	1.9	4.2	0.5	0.6	**0.0**
		@ 0.5	94.5	83.8	**0.1**	5.4	9.5	23.6	0.6	0.6	0.9

NEURAL INFORMATION PROCESSING SYSTEMS

Summary

- **GoF tests are simple yet powerful tool for watermark detection in LLM.**
  - Ability of capturing distribution level differences
- **Test repetition is a key factor that enhances GoF performance at low temperatures.**
- **Some future directions:**
  - A deeper theoretical understanding of GoF test in detection.
  - Adaptive detection strategies that dynamically select among GoF tests.
  - Potential risks of GoF tests, like type I error, lack of non-asymptotic distribution, and how to handle these.

📄 **Our paper is also on arXiv:** https://arxiv.org/abs/2510.03944
💻 **Code is available on Github:** https://github.com/hwq0726/GoF-for-Watermark-Detection
✉️ **Feel free to contact us:** weiqingh@sas.upenn.edu

Penn
University of Pennsylvania