

Training-free LLM-generated Text Detection by Mining Token Probability Sequences

Yihuai Xu^{1,4}, Yongwei Wang^{1,*}, Yifei Bi^{1,2}, Huangsen Cao¹, Zhouhan Lin³, Yu Zhao¹, Fei Wu¹

¹Zhejiang University,

²Georgia Institute of Technology,

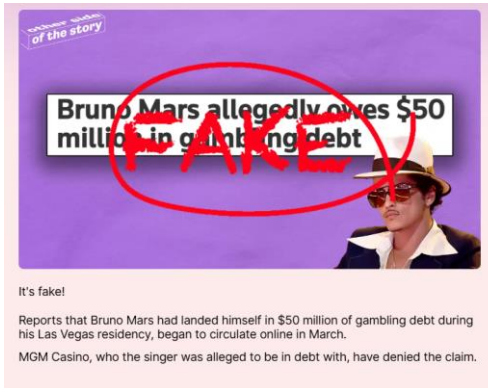
³Shanghai Jiao Tong University,

⁴Zhejiang Gongshang University

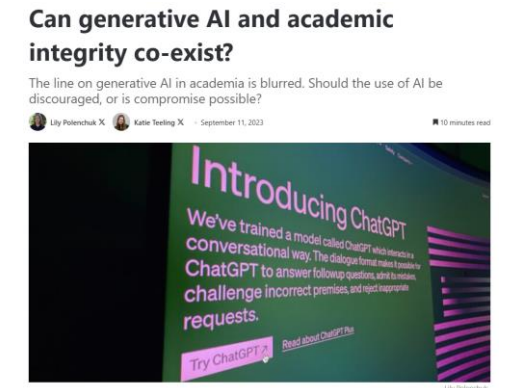
Background

The increasing sophistication of LLMs has raised serious concerns about their **potential misuse**.

Fake News

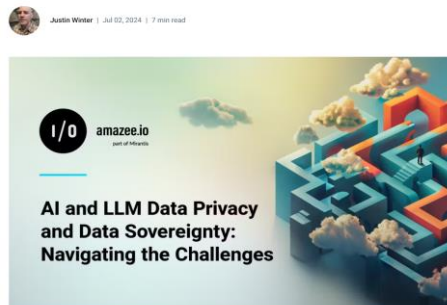


Academic Dishonesty

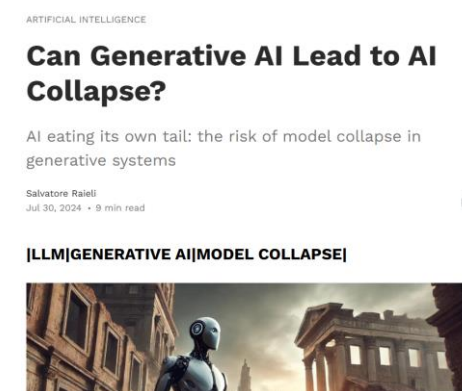


Data Privacy

AI and LLM Data Privacy and Data Sovereignty: Navigating the Challenges



Model Collapse



How to detect LLM-generated text?



There are **three** main methods to detect LLM-generated text:

1. Training-free (**main focus**) :

- Sample-based : Likelihood, Rank, Entropy, DetectLRR.
- Distribution-based : DetectGPT, DNA-GPT, Fast-DetectGPT.

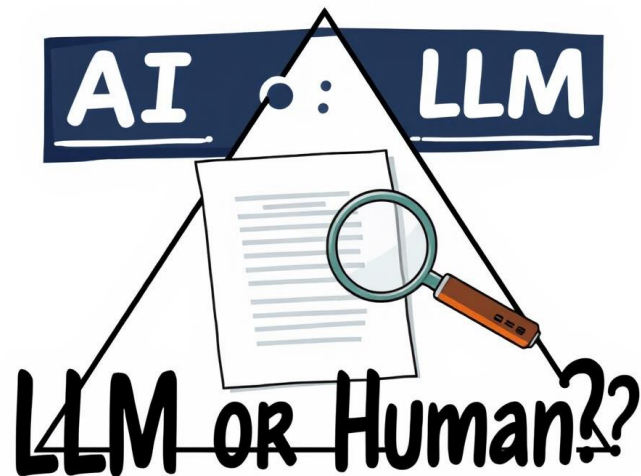
Advantages of training-free methods:

- decent performance
- simple deployment
- strong generalization


2. Training-based :

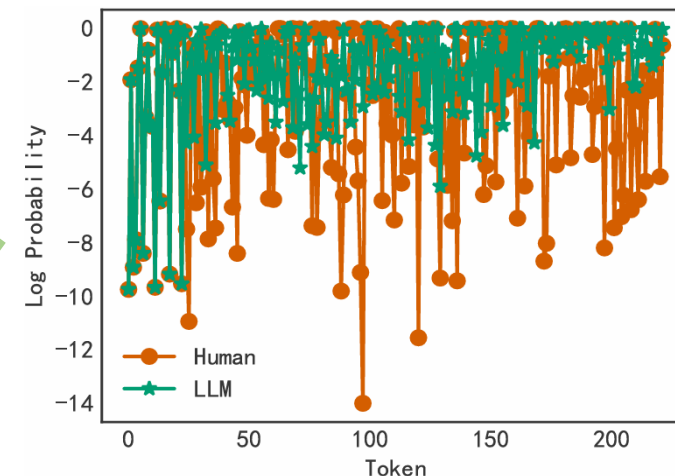
- RoBERTa, SeqXGPT, RADAR, MPU, ReMoDetect.
- Some *commercial detectors* like GPTZero.

3. Watermarking

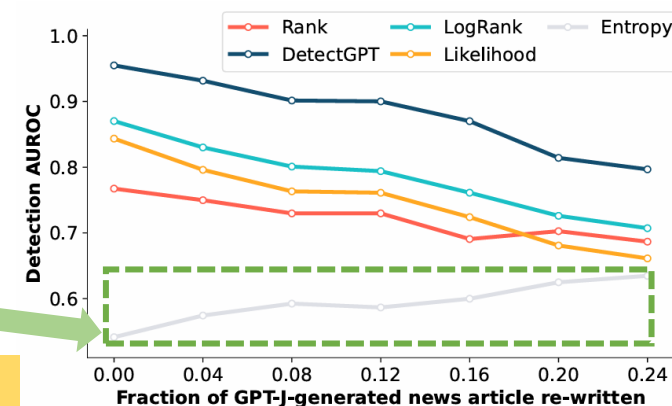


Our observations:

1. The token probability sequences (TPS) of human-written text and LLM-generated text exhibit significant differences in their **fluctuations**. 
2. Mining the **local dynamic patterns** of TPS is evidently promising for enhancing detection performance (previous detectors based on **global static patterns** neglected it).
3. According to DetectGPT, the **entropy** method demonstrates a particularly distinctive performance in countering paraphrasing attacks.



An illustration example comparing the fluctuations of *token probability sequence* (TPS) between human-written and LLM-generated texts (with OPT-2.7).



From the original paper of DetectGPT (Figure 5)

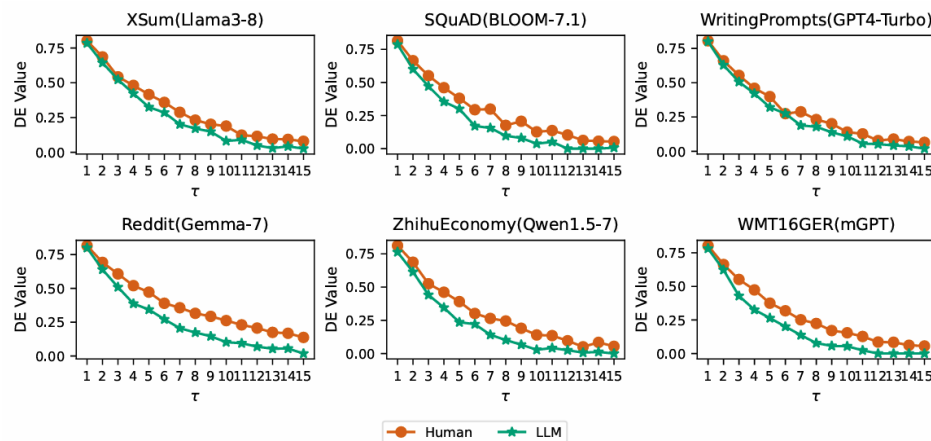
Dynamic Complexity Modeling Using Entropy-Based Time Series Analysis !

A novel training-free detector : Lastde

- Applying multi-scale diversity entropy (MDE, a 1-dim *time series analysis* method) to extract the temporal dynamics of TPS and further developing the *local* statistic called **Agg-MDE**.
- Combining with the classic *global* statistic **Log-Likelihood**.

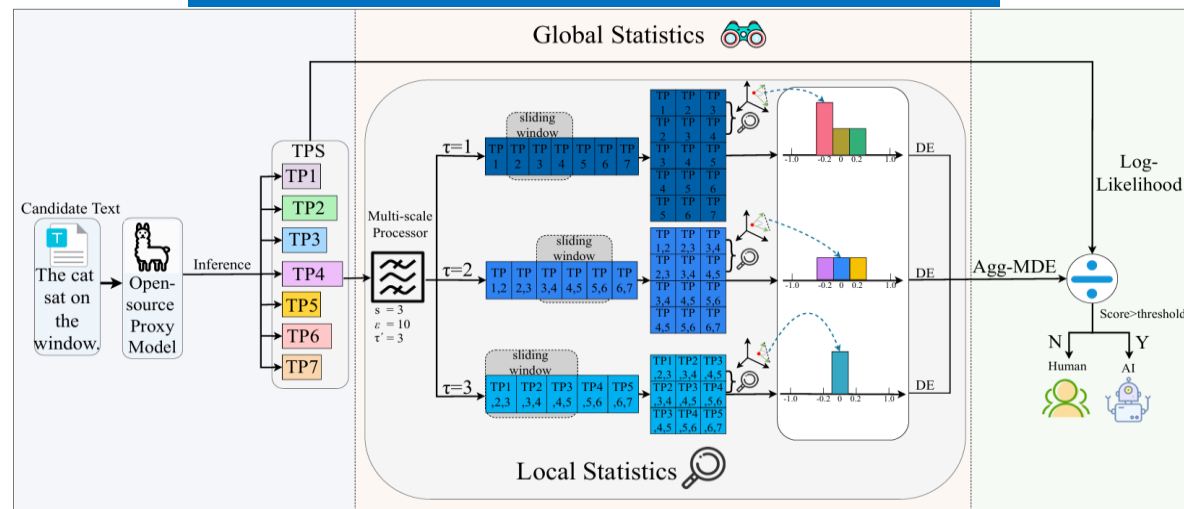
Key formulae

$$DE(s, \varepsilon, \tau) = -\frac{1}{\ln \varepsilon} \sum_{i=1}^{\varepsilon} P_i^{(\tau)} \ln P_i^{(\tau)}, \quad P_i^{(\tau)} > 0.$$

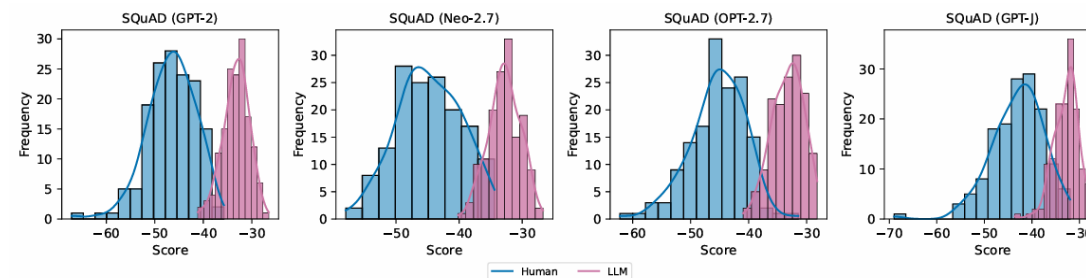


$$\text{Lastde}(\mathbf{t}, \theta) = \frac{\frac{1}{n} \sum_{i=1}^n \log p_{\theta}(t_i | t_{<i})}{\text{Agg}((DE(s, \varepsilon, 1), \dots, DE(s, \varepsilon, \tau')))},$$

Overview of the Lastde detection framework

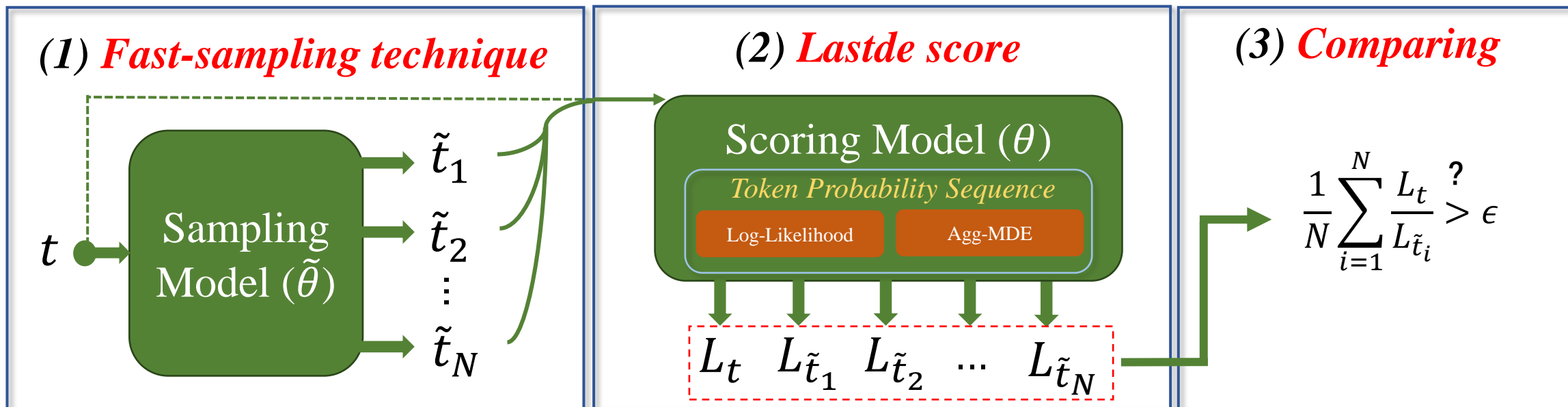


Lastde score

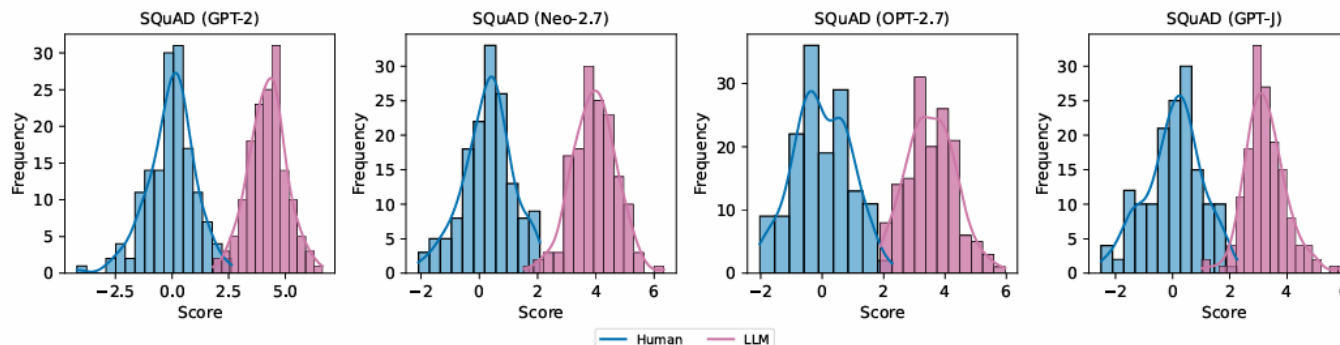


Fast-sampling variant: Lastde++

Following the *fast-sampling technique* of Fast-DetectGPT, we can obtain a more powerful variant : **Lastde++** .



**Lastde++
score**



Main results

- 2 conventional scenarios :

- White box
- Black box

- 4 different domains :

- XSum
- SQuAD
- WritingPrompts
- Reddit

- 12 open-source models :

- GPT-2
- ...
- Phi2-2.7

- 3 closed-source models :

- GPT-4-Turbo
- GPT-4o
- Claude3-haiku

White box

Methods/Models	GPT-2	Neo-2.7	OPT-2.7	GPT-J	Llama-13	Llama2-13	Llama3-8	OPT-13	BLOOM-7.1	Falcon-7	Gemma-7	Phi2-2.7	Avg.
Sample-based Methods													
Likelihood	91.65	89.40	88.08	84.95	63.66	65.36	98.35	84.45	88.00	76.78	70.14	89.67	82.54
LogRank	94.31	92.87	90.99	88.68	68.87	70.27	99.04	87.74	92.42	81.32	74.81	92.13	86.12
Entropy	52.15	51.72	50.46	54.31	64.18	61.05	23.30	54.30	62.67	59.33	66.47	44.09	53.67
DetectLRR	96.67	96.07	93.13	92.24	81.40	80.89	98.94	91.03	96.35	87.45	81.36	94.10	90.80
Lastde	98.41	98.64	98.15	97.24	88.98	88.40	99.71	96.47	99.35	95.49	91.85	96.99	95.89
(Diff)	1.74	2.57	5.02	5.00	8.58	7.51	0.77	5.44	3.00	8.04	10.5	2.89	5.09
Distribution-based Methods													
DetectGPT	93.43	90.40	90.36	83.82	63.78	65.39	70.13	85.05	89.28	77.98	68.96	89.55	80.68
DetectNPR	95.77	94.77	93.24	88.86	68.60	69.83	95.55	89.78	94.95	83.06	74.74	93.06	86.85
DNA-GPT	89.92	86.80	86.79	82.21	62.28	64.46	98.07	82.51	86.74	74.04	63.63	88.00	80.45
Fast-DetectGPT	99.57	99.49	98.78	98.95	93.45	93.34	99.91	98.07	99.53	97.74	96.90	98.10	97.82
Lastde++	99.76	99.87	99.46	99.52	96.58	96.67	99.82	98.77	99.84	98.76	98.40	98.76	98.85
(Diff)	0.19	0.38	0.68	0.57	3.13	3.33	-0.09	0.70	0.31	1.02	1.50	0.66	1.03

Black box

Methods/Models	GPT-2	Neo-2.7	OPT-2.7	Llama-13	Llama2-13	Llama3-8	OPT-13	BLOOM-7.1	Falcon-7	Gemma-7	Phi2-2.7	GPT-4-Turbo	Avg.
Sample-based Methods													
Likelihood	65.88	67.09	67.40	65.75	68.61	99.60	68.80	61.80	67.42	69.90	73.93	79.69	71.32
LogRank	70.38	71.17	72.35	70.28	72.67	99.69	73.01	67.51	71.66	72.17	77.99	79.24	74.84
Entropy	61.48	58.65	54.55	49.14	45.18	14.43	53.09	60.84	50.55	48.01	46.58	35.09	48.13
DetectLRR	79.30	79.19	81.25	78.51	78.94	97.35	80.27	79.57	79.87	73.47	83.79	73.85	80.45
Lastde	89.17	90.24	89.70	80.71	79.90	99.67	90.01	88.94	84.36	79.61	88.32	81.33	86.38
(Diff)	9.87	11.1	8.45	2.20	0.96	2.32	9.74	9.37	4.49	6.14	4.53	7.48	6.38
Distribution-based Methods													
DetectGPT	67.56	69.28	72.03	66.12	67.96	82.90	73.89	61.83	68.69	66.55	72.76	81.73	70.94
DetectNPR	68.07	68.41	73.06	67.83	70.60	96.75	75.13	63.00	70.42	65.72	74.08	79.94	72.75
DNA-GPT	64.15	62.63	63.64	60.77	66.71	99.47	65.75	62.01	65.08	62.59	72.02	70.75	67.97
Fast-DetectGPT	89.82	88.75	86.52	77.58	77.62	99.43	86.16	84.55	81.42	81.49	86.67	88.18	85.68
Lastde++	94.93	95.28	94.13	85.00	85.80	99.03	93.37	92.22	89.49	87.58	92.67	88.21	91.47
(Diff)	5.11	6.53	7.62	7.42	8.18	-0.40	7.21	7.67	8.07	6.09	6.00	0.03	5.80

Black box

Methods ↓	SourceModels → Datasets →	GPT-4-Turbo				GPT-4o				Claude-3-haiku			
		XSum	WritingPrompts	Reddit	Avg.	XSum	WritingPrompts	Reddit	Avg.	XSum	WritingPrompts	Reddit	Avg.
Likelihood		60.44	81.48	97.15	79.69	75.42	84.90	97.74	86.02	96.84	98.38	99.92	98.38
LogRank		61.52	79.03	97.16	79.24	73.85	82.32	97.74	84.64	97.09	98.71	99.96	98.59
Entropy		61.24	35.56	08.48	35.09	47.50	31.60	09.74	29.61	38.90	17.69	06.56	21.05
DetectLRR		61.71	66.75	93.10	73.85	62.87	69.06	93.75	75.23	95.78	97.96	99.56	97.77
Lastde(Std)		64.16	83.09	96.74	81.33	73.87	86.20	97.74	85.94	97.44	99.40	99.92	98.92
Aggregation function													
Fast-DetectGPT		80.79	89.88	93.87	88.18	86.87	93.77	97.93	92.86	99.93	99.99	99.96	99.96
Lastde++(2-Norm)		76.91	87.39	93.61	85.97	85.74	92.96	97.52	92.07	99.95	99.99	99.96	99.97
Lastde++(Range)		82.67	86.37	91.72	86.92	85.96	91.34	96.57	91.29	99.84	99.99	99.95	99.93
Lastde++(Std)		83.12	88.50	93.00	88.21	86.47	93.41	96.98	92.29	99.92	99.96	99.89	99.92
Lastde++(ExpRange)		82.40	89.02	93.70	88.37	87.42	93.60	97.77	92.93	99.96	100.00	99.97	99.98
Lastde++(ExpStd)		81.55	89.81	93.99	88.45	87.24	94.24	97.94	93.14	99.97	99.99	99.95	99.97

Main results

- **More datasets** : ReMoDetect and Fast-DetectGPT benchmark datasets.
 - 4 domain : XSum, SQuAD, WritingPrompts and PubMedQA.
 - 6 source models (parameters $\geq 20B$) : NeoX-20B, Llama3-70B, GPT-3.5-Turbo, GPT-4, GPT-4-Turbo and Claude-3.
- **More baselines** : Binoculars and GPTZero (commercial, 2024-11-11 base version).

Models	Domains	Likelihood	LogRank	DetectLRR	Fast-DetectGPT	Binoculars	GPTZero	Lastde	Lastde++
NeoX-20B	XSum	68.1	69.9	70.6	84.7	54.3	60.1	78.8	88.4
	SQuAD	64.8	69.3	77.6	85.9	56.6	52.6	86.2	91.8
	WritingPrompts	87.7	89.7	90.7	96.7	66.1	72.4	94.7	97.4
Llama3-70B	XSum	96.9	97.4	94.9	99.9	98.7	100	97.2	99.9
	WritingPrompts	98.2	97.9	93.8	99.9	100	99.8	98.2	99.9
	PubMedQA	84.8	83.5	71.1	90.5	88.7	90.1	85.6	90.7
GPT-4-Turbo	XSum	87.8	89.1	87.2	98.3	94.9	100	88.9	98.5
	WritingPrompts	98.2	98.0	94.4	99.9	99.3	100	98.2	99.8
	PubMedQA	85.8	85.1	74.5	87.8	90.0	87.2	86.4	87.8
GPT-4	XSum	73.7	73.7	69.8	91.6	78.7	98.2	74.1	91.6
	WritingPrompts	87.6	85.5	73.4	97.6	90.7	82.6	87.3	97.4
	PubMedQA	79.6	79.0	69.3	83.7	86.7	84.8	80.4	83.8
GPT-3.5-Turbo	XSum	93.8	94.1	90.9	99.2	99.7	99.5	94.2	99.1
	WritingPrompts	98.1	97.7	93.0	99.7	99.3	92.9	98.2	99.6
	PubMedQA	87.2	86.6	75.8	90.4	93.0	88.0	88.1	90.4
Claude-3	XSum	91.8	92.6	89.8	96.4	94.0	99.9	92.4	96.4
	WritingPrompts	97.0	96.4	88.8	96.1	96.0	99.1	97.0	96.0
	PubMedQA	85.5	84.9	74.8	88.0	88.0	88.0	86.2	87.9
Avg.		87.0	87.2	82.2	93.7	87.5	88.6	89.6	94.2

Complex scenarios

- **2 Complex scenarios** : paraphrasing attack and cross-lingual.
- **3 different languages** : English, German and Chinese.

Robustness against paraphrasing attack

White-box

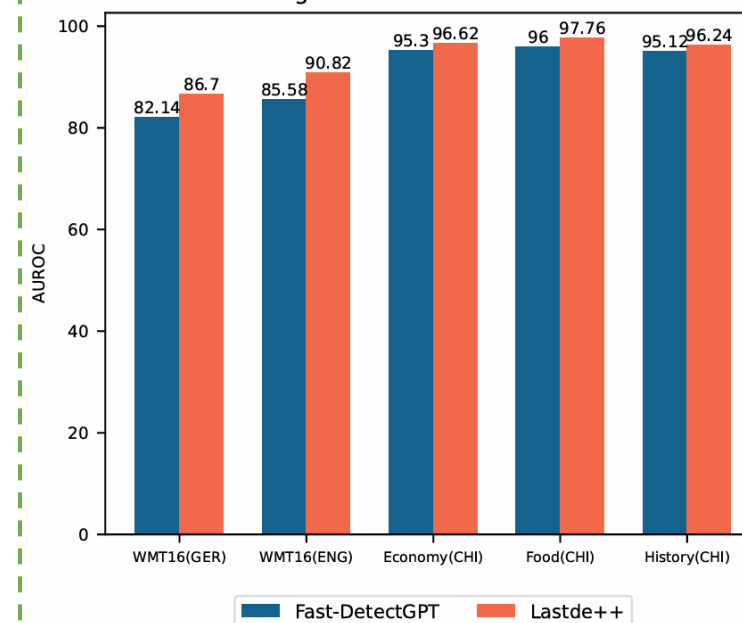
Methods/Datasets	XSum(Original)	XSum(Paraphrased)	WritingPrompts(Original)	WritingPrompts(Paraphrased)	Reddit(Original)	Reddit(Paraphrased)
<i>Llama-13</i>						
Fast-DetectGPT	94.80	81.12 (↓ 13.68)	98.35	95.01 (↓ 3.34)	96.84	91.22 (↓ 5.62)
Lastde++	97.27	85.65 (↓ 11.62)	99.56	97.69 (↓ 1.87)	98.48	95.15 (↓ 3.33)
<i>OPT-13</i>						
Fast-DetectGPT	95.19	85.43 (↓ 9.76)	99.40	97.46 (↓ 1.94)	98.81	96.61 (↓ 2.20)
Lastde++	96.94	88.30 (↓ 8.64)	99.40	97.43 (↓ 1.97)	99.43	98.00 (↓ 1.43)
<i>GPT-J</i>						
Fast-DetectGPT	98.12	89.88 (↓ 8.24)	99.22	96.40 (↓ 2.82)	98.58	95.97 (↓ 2.61)
Lastde++	99.04	93.87 (↓ 5.17)	99.75	98.41 (↓ 1.34)	99.38	97.85 (↓ 1.53)

Black-box

Methods/Datasets	XSum(Original)	XSum(Paraphrased)	WritingPrompts(Original)	WritingPrompts(Paraphrased)	Reddit(Original)	Reddit(Paraphrased)
<i>Llama-13</i>						
Fast-DetectGPT	64.79	53.91 (↓ 10.88)	88.26	82.09 (↓ 6.17)	79.68	74.05 (↓ 5.63)
Lastde++	72.84	61.27 (↓ 11.57)	94.46	90.05 (↓ 4.41)	87.70	81.68 (↓ 6.02)
<i>OPT-13</i>						
Fast-DetectGPT	84.27	72.63 (↓ 11.64)	89.64	88.44 (↓ 1.20)	84.57	83.73 (↓ 0.84)
Lastde++	91.76	80.58 (↓ 11.18)	95.25	94.24 (↓ 1.01)	93.09	91.27 (↓ 1.82)
<i>GPT-4-Turbo</i>						
Fast-DetectGPT	80.79	74.96 (↓ 5.83)	89.88	80.93 (↓ 8.95)	93.87	91.35 (↓ 2.52)
Lastde++	83.12	78.44 (↓ 4.68)	88.50	81.32 (↓ 7.18)	93.00	90.74 (↓ 2.26)

Cross-lingual robustness

Average AUROC for Both Scenarios



Complex scenarios

- 3 different metrics :

- AUROC
- TPR at 5% FPR
- Accuracy at 5% FPR

- 2 complex scenarios :

- Datasets for human-style imitation
- Datasets for LLM-LLM mixture text

- 3 different mixing ratios :

- 50% + 50% (2 source models)
- 80% + 20% (2 source models)
- 25% each one (4 source models)

The average of WritingPrompts (GPT-4-turbo-0409, GPT-4o-08-06)

	Likelihood	LogRank	DetectLRR	Lastde	Fast-DetectGPT	Lastde++
AUROC	87.37	85.60	74.94	87.68	96.49	96.49
TPR at 5% FPR	30.00	30.67	24.34	29.33	82.67	85.36
Accuracy at 5% FPR	62.50	62.83	59.67	63.34	88.84	90.17

Three metrics on LLM-LLM mixture text datasets

Table 17: AUROC on LLM-LLM mixture text

Mixing ratio	Source Models	Likelihood	LogRank	DetectLRR	Lastde	Fast-DetectGPT	Lastde++
50%+50%	(Llama2-13B, OPT-13B)	81.48	84.81	87.99	91.88	85.21	92.48
	(BLOOM-7B, Falcon-7B)	78.80	82.56	87.47	94.95	88.59	95.65
80%+20%	(Llama2-13B, OPT-13B)	80.44	82.97	85.45	89.13	86.73	92.64
	(BLOOM-7B, Falcon-7B)	76.98	81.52	52.91	96.38	92.05	96.57
25% each	(Llama2-13B, OPT-13B, BLOOM-7B, Falcon-7B)	82.33	86.34	88.02	94.44	86.02	93.32

Table 18: TPR at 5% FPR on LLM-LLM mixture text

Mixing ratio	Source Models	Likelihood	LogRank	DetectLRR	Lastde	Fast-DetectGPT	Lastde++
50%+50%	(Llama2-13B, OPT-13B)	29.33	36.00	40.00	65.33	46.67	68.00
	(BLOOM-7B, Falcon-7B)	16.00	30.00	34.67	76.00	39.33	70.67
80%+20%	(Llama2-13B, OPT-13B)	24.67	28.67	38.67	58.67	41.33	69.33
	(BLOOM-7B, Falcon-7B)	16.67	27.67	35.33	59.67	49.33	83.33
25% each	(Llama2-13B, OPT-13B, BLOOM-7B, Falcon-7B)	22.00	40.67	46.67	81.33	40.67	73.33

Table 19: Accuracy at 5% FPR on LLM-LLM mixture text

Mixing ratio	Source Models	Likelihood	LogRank	DetectLRR	Lastde	Fast-DetectGPT	Lastde++
50%+50%	(Llama2-13B, OPT-13B)	62.17	65.50	67.50	80.17	70.83	81.50
	(BLOOM-7B, Falcon-7B)	55.55	62.50	64.83	85.50	67.17	82.83
80%+20%	(Llama2-13B, OPT-13B)	59.83	61.83	66.83	76.83	68.17	82.17
	(BLOOM-7B, Falcon-7B)	55.83	58.83	69.83	87.83	77.17	89.17
25% each	(Llama2-13B, OPT-13B, BLOOM-7B, Falcon-7B)	58.50	67.83	70.50	88.17	67.83	84.17

Main contributions :

- Lastde (and Lastde++), a novel and effective method for LLM-generated text detection.
- Mining token probability sequences (*temporal dynamics*) from a time series analysis viewpoint.
- Integrating *local* and *global statistics* for more robust detection.

Advantages :

- state-of-the-art detection performance (surpassing advanced training-free baseline methods).
- comparable or lower computational costs (comparing to DetectGPT, DNA-GPT, and Fast-DetectGPT).
- superior robustness (including cross-domain, cross-model, cross-lingual tasks, and paraphrasing attacks).