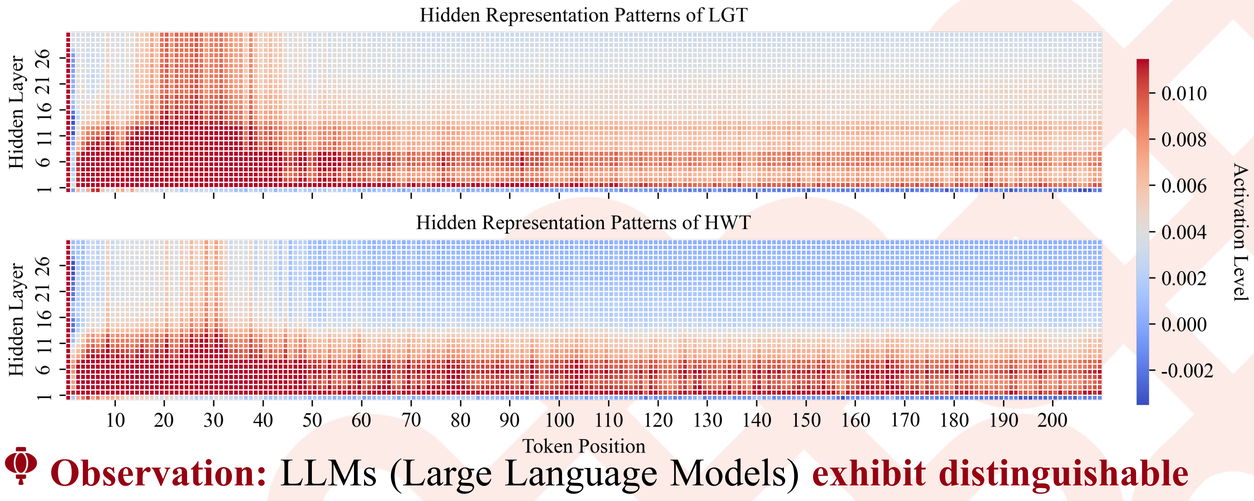
RepreGuard: Detecting LLM-Generated Text by Revealing Hidden Representation Patterns

Xin Chen[♤♡], Junchao Wu[♤], Shu Yang[♤], Runzhe Zhan[♤], Zeyu Wu[♤], Ziyang Luo[◊]
Di Wang[♤], Min Yang[♡], Lidia S. Chao[♤], Derek F. Wong[♤]

[♤]NLP2CT Lab, Department of Computer and Information Science, University of Macau
[♡]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences; [♤]KAUST, [◊]HKBU

Background and Motivation

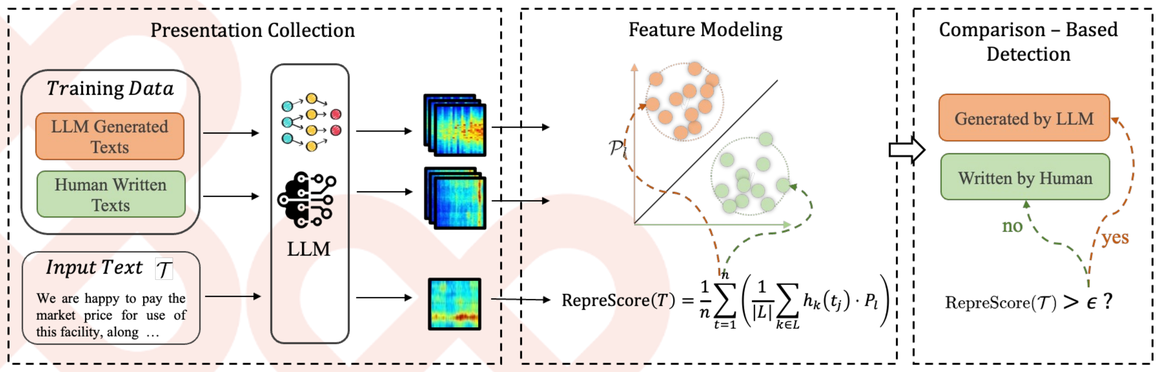
- **Background:** Academic dishonesty, the spread of fake news, user mistrust in digital content, data pollution in web-sourced data…
- **Motivation:** Current methods **rely on large amounts of training data** or single dimension with **poor robustness** across models and domains.



- **Observation:** LLMs (Large Language Models) **exhibit distinguishable hidden representation patterns when processing human-written texts (HWT) and LLM-generated texts (LGT),** offering a promising method for achieving more robust LGT detection.

Methodology

- **Hypothesis:** The premise of RepreGuard is that LLMs perceive the statistical patterns of LGT and HWT differently. Their hidden representation patterns exhibit differences when observing and processing these two types of text.



- **Representation Collection:** We first introduce a small training set, formalized as $\{(\mathcal{T}\_LGT\hat{}i, \mathcal{T}\_HWT\hat{}i) | i \in [1, \mathcal{N}], \ell \in [1, \mathcal{L}] \backslash \}\}$, where $\mathcal{N}$ is the number of LGT and HWT pairs in the dataset. For each token t_j in $\mathcal{T}$, we collect the neural activation of the model $\mathcal{M}$ at each layer ℓ, represented as $h\_j\hat{}\ell \in \mathbb{R}\hat{}d \in \mathcal{R}$, where d is the dimension of the model's hidden state. Based on this, the complete activation of the text T in model $\mathcal{M}$ can be formalized:

$$\mathcal{A}(\mathcal{T}) = \left\{h_j^{\ell} \mid j \in [1, \mathcal{N}], \ell \in [1, \mathcal{L}]\right\}$$

- **Feature Modeling:** We perform PCA on difference in activation patterns between LGT and HWT for each text pair:

$$\mathcal{P}_\ell = \left\{PCA\left(\mathcal{A}^{\ell}\left(\mathcal{T}_{LGT}^{i}\right) - \mathcal{A}^{\ell}\left(\mathcal{T}_{HWT}^{i}\right)\right)\right.$$

- **Comparison-based Detection:** We project the activations of each token t_j in text $\mathcal{T}$ onto the probing vector across all layers $\mathcal{L}$, defining this as:

$$\text{RepreScore}(T) = \frac{1}{\mathcal{N}}\sum_{t=1}^{n}\left(\frac{1}{|\mathcal{L}|}\sum_{k\in\mathcal{L}} h_k(t_j)\cdot\mathcal{P}_\ell\right)$$

Main Result

Best Average AUROC (AUR.) And TPR@0.01 (TPR.) Performance on In-domain and Out-of-domain[1]

Test→ / Train↓	Detector↓ / Metrics→	ChatGPT *AUR.*	ChatGPT *TPR.*	Llama-2-70b *AUR.*	Llama-2-70b *TPR.*	Google-PaLM *AUR.*	Google-PaLM *TPR.*	Claude-instant *AUR.*	Claude-instant *TPR.*	Avg. *AUR.*	Avg. *TPR.*
ChatGPT	Roberta	98.38$_{\pm 0.32}$	90.52$_{\pm 1.93}$	81.71$_{\pm 2.27}$	54.64$_{\pm 16.57}$	74.56$_{\pm 0.89}$	20.20$_{\pm 12.87}$	66.74$_{\pm 1.49}$	22.36$_{\pm 13.26}$	80.35$_{\pm 1.24}$	46.93$_{\pm 11.16}$
	LRR	92.61$_{\pm 0.39}$	26.20$_{\pm 0.00}$	95.84$_{\pm 0.34}$	86.20$_{\pm 0.00}$	81.98$_{\pm 0.14}$	39.80$_{\pm 0.00}$	57.82$_{\pm 0.80}$	0.10$_{\pm 0.00}$	82.06$_{\pm 0.42}$	38.07$_{\pm 0.00}$
	DetectGPT	54.87$_{\pm 0.25}$	0.11$_{\pm 0.14}$	59.21$_{\pm 0.53}$	0.66$_{\pm 1.77}$	55.29$_{\pm 0.37}$	0.08$_{\pm 0.06}$	55.92$_{\pm 0.61}$	0.00$_{\pm 0.00}$	56.32$_{\pm 0.44}$	0.21$_{\pm 0.49}$
	Fast-Detect.	75.65$_{\pm 0.28}$	11.60$_{\pm 0.00}$	85.49$_{\pm 0.31}$	2.50$_{\pm 0.00}$	80.36$_{\pm 0.63}$	17.70$_{\pm 0.00}$	47.29$_{\pm 0.44}$	0.00$_{\pm 0.00}$	72.20$_{\pm 0.42}$	7.95$_{\pm 0.00}$
	Str-Detect.	52.05$_{\pm 0.00}$	0.01$_{\pm 0.00}$	52.30$_{\pm 0.00}$	0.01$_{\pm 0.00}$	56.05$_{\pm 0.00}$	0.01$_{\pm 0.00}$	57.50$_{\pm 0.00}$	0.01$_{\pm 0.00}$	54.47$_{\pm 0.00}$	0.01$_{\pm 0.00}$
	Binoculars	97.36$_{\pm 0.52}$	40.70$_{\pm 0.00}$	99.45$_{\pm 0.44}$	98.70$_{\pm 0.00}$	**98.03$_{\pm 0.34}$**	**89.80$_{\pm 0.00}$**	61.86$_{\pm 3.64}$	3.40$_{\pm 0.00}$	89.18$_{\pm 1.24}$	58.15$_{\pm 0.00}$
	RepreGuard	**99.84$_{\pm 0.12}$**	**100.00$_{\pm 0.00}$**	**99.55$_{\pm 0.12}$**	**99.26$_{\pm 0.11}$**	88.67$_{\pm 0.65}$	64.66$_{\pm 1.26}$	**85.00$_{\pm 1.40}$**	**56.12$_{\pm 2.20}$**	**93.26$_{\pm 0.57}$**	**80.01$_{\pm 0.89}$**
Llama-2-70b	Roberta	95.49$_{\pm 0.96}$	71.16$_{\pm 4.59}$	94.21$_{\pm 2.06}$	63.66$_{\pm 10.40}$	86.00$_{\pm 2.17}$	43.60$_{\pm 2.56}$	76.98$_{\pm 4.04}$	22.88$_{\pm 5.13}$	88.17$_{\pm 2.31}$	50.32$_{\pm 5.67}$
	LRR	91.88$_{\pm 1.05}$	26.20$_{\pm 0.00}$	96.47$_{\pm 0.52}$	86.20$_{\pm 0.00}$	81.65$_{\pm 0.55}$	39.80$_{\pm 0.00}$	56.20$_{\pm 1.77}$	0.10$_{\pm 0.00}$	81.55$_{\pm 0.97}$	38.07$_{\pm 0.00}$
	DetectGPT	54.33$_{\pm 0.81}$	0.51$_{\pm 0.80}$	59.36$_{\pm 0.86}$	1.28$_{\pm 2.18}$	55.02$_{\pm 0.98}$	0.02$_{\pm 0.06}$	56.11$_{\pm 0.24}$	0.00$_{\pm 0.00}$	56.20$_{\pm 0.72}$	0.45$_{\pm 0.76}$
	Fast-Detect.	94.08$_{\pm 1.26}$	71.90$_{\pm 0.00}$	98.76$_{\pm 0.05}$	94.20$_{\pm 0.00}$	92.39$_{\pm 0.28}$	81.80$_{\pm 0.00}$	51.45$_{\pm 0.62}$	0.40$_{\pm 0.00}$	84.17$_{\pm 0.55}$	62.08$_{\pm 0.00}$
	Str-Detect.	52.05$_{\pm 0.00}$	0.01$_{\pm 0.00}$	52.30$_{\pm 0.00}$	0.01$_{\pm 0.00}$	56.05$_{\pm 0.00}$	0.01$_{\pm 0.00}$	57.50$_{\pm 0.00}$	0.01$_{\pm 0.00}$	54.47$_{\pm 0.00}$	0.01$_{\pm 0.00}$
	Binoculars	97.94$_{\pm 0.74}$	85.90$_{\pm 0.00}$	**99.63$_{\pm 0.06}$**	98.70$_{\pm 0.00}$	**97.23$_{\pm 0.93}$**	**89.80$_{\pm 0.00}$**	57.49$_{\pm 3.57}$	3.40$_{\pm 0.00}$	88.07$_{\pm 1.32}$	69.45$_{\pm 0.00}$
	RepreGuard	**99.54$_{\pm 0.08}$**	**99.38$_{\pm 0.14}$**	99.38$_{\pm 0.18}$	**98.84$_{\pm 0.11}$**	88.84$_{\pm 1.28}$	77.08$_{\pm 0.45}$	**84.08$_{\pm 3.52}$**	**60.66$_{\pm 1.66}$**	**92.96$_{\pm 1.27}$**	**83.99$_{\pm 0.59}$**
Google-PaLM	Roberta	88.72$_{\pm 1.35}$	40.94$_{\pm 5.18}$	85.36$_{\pm 5.00}$	32.70$_{\pm 6.15}$	82.09$_{\pm 3.99}$	36.52$_{\pm 4.64}$	72.98$_{\pm 4.00}$	21.54$_{\pm 8.64}$	82.29$_{\pm 3.58}$	32.92$_{\pm 6.15}$
	LRR	91.40$_{\pm 2.01}$	26.20$_{\pm 0.00}$	92.84$_{\pm 2.78}$	86.20$_{\pm 0.00}$	83.13$_{\pm 1.53}$	39.80$_{\pm 0.00}$	61.11$_{\pm 2.05}$	0.10$_{\pm 0.00}$	82.12$_{\pm 2.09}$	38.07$_{\pm 0.00}$
	DetectGPT	54.04$_{\pm 0.32}$	0.00$_{\pm 0.00}$	59.21$_{\pm 0.22}$	0.00$_{\pm 0.00}$	55.30$_{\pm 0.35}$	0.02$_{\pm 0.06}$	55.53$_{\pm 0.46}$	0.96$_{\pm 1.63}$	56.02$_{\pm 0.34}$	0.24$_{\pm 0.42}$
	Fast-Detect.	74.62$_{\pm 0.99}$	11.60$_{\pm 0.00}$	86.47$_{\pm 0.50}$	2.50$_{\pm 0.00}$	80.64$_{\pm 0.22}$	17.70$_{\pm 0.00}$	48.46$_{\pm 0.47}$	0.00$_{\pm 0.00}$	72.55$_{\pm 0.54}$	7.95$_{\pm 0.00}$
	Str-Detect.	52.05$_{\pm 0.00}$	0.01$_{\pm 0.00}$	52.30$_{\pm 0.00}$	0.01$_{\pm 0.00}$	56.05$_{\pm 0.00}$	0.01$_{\pm 0.00}$	57.50$_{\pm 0.00}$	0.01$_{\pm 0.00}$	54.47$_{\pm 0.00}$	0.01$_{\pm 0.00}$
	Binoculars	**98.56$_{\pm 0.35}$**	85.90$_{\pm 0.00}$	**99.58$_{\pm 0.20}$**	98.70$_{\pm 0.00}$	**97.98$_{\pm 0.38}$**	**89.80$_{\pm 0.00}$**	61.15$_{\pm 2.49}$	3.40$_{\pm 0.00}$	89.32$_{\pm 0.86}$	69.45$_{\pm 0.00}$
	RepreGuard	98.39$_{\pm 0.31}$	**99.36$_{\pm 0.07}$**	98.53$_{\pm 0.28}$	**99.16$_{\pm 0.07}$**	93.36$_{\pm 0.27}$	79.88$_{\pm 3.43}$	**90.57$_{\pm 1.06}$**	**56.90$_{\pm 2.20}$**	**95.21$_{\pm 0.48}$**	**83.82$_{\pm 1.44}$**
Claude-instant	Roberta	88.63$_{\pm 1.34}$	28.50$_{\pm 11.74}$	77.45$_{\pm 4.73}$	23.70$_{\pm 10.19}$	74.90$_{\pm 1.76}$	14.58$_{\pm 13.14}$	88.07$_{\pm 3.89}$	36.56$_{\pm 4.83}$	82.26$_{\pm 2.93}$	25.84$_{\pm 9.98}$
	LRR	86.33$_{\pm 4.23}$	26.20$_{\pm 0.00}$	87.46$_{\pm 4.43}$	86.20$_{\pm 0.00}$	81.19$_{\pm 3.34}$	39.80$_{\pm 0.00}$	63.74$_{\pm 1.07}$	0.10$_{\pm 0.00}$	79.68$_{\pm 3.27}$	38.07$_{\pm 0.00}$
	DetectGPT	53.95$_{\pm 0.84}$	0.02$_{\pm 0.06}$	57.90$_{\pm 1.09}$	0.00$_{\pm 0.00}$	54.33$_{\pm 1.13}$	0.06$_{\pm 0.07}$	55.97$_{\pm 0.38}$	0.00$_{\pm 0.00}$	55.54$_{\pm 0.86}$	0.02$_{\pm 0.03}$
	Fast-Detect.	81.27$_{\pm 5.40}$	71.90$_{\pm 0.00}$	82.35$_{\pm 4.48}$	94.20$_{\pm 0.00}$	80.85$_{\pm 4.70}$	81.80$_{\pm 0.00}$	61.35$_{\pm 0.35}$	0.40$_{\pm 0.00}$	76.46$_{\pm 3.73}$	62.08$_{\pm 0.00}$
	Str-Detect.	52.05$_{\pm 0.00}$	0.01$_{\pm 0.00}$	52.30$_{\pm 0.00}$	0.01$_{\pm 0.00}$	56.05$_{\pm 0.00}$	0.01$_{\pm 0.00}$	57.50$_{\pm 0.00}$	0.01$_{\pm 0.00}$	54.47$_{\pm 0.00}$	0.01$_{\pm 0.00}$
	Binoculars	92.36$_{\pm 1.95}$	85.90$_{\pm 0.00}$	93.18$_{\pm 1.01}$	**98.70$_{\pm 0.00}$**	**92.83$_{\pm 1.18}$**	**89.80$_{\pm 0.00}$**	73.92$_{\pm 0.17}$	3.40$_{\pm 0.00}$	88.07$_{\pm 1.08}$	69.45$_{\pm 0.00}$
	RepreGuard	**97.20$_{\pm 1.39}$**	**96.92$_{\pm 0.49}$**	**97.51$_{\pm 0.99}$**	97.16$_{\pm 0.40}$	87.77$_{\pm 2.37}$	63.04$_{\pm 0.73}$	**92.76$_{\pm 0.49}$**	**56.22$_{\pm 5.22}$**	**93.81$_{\pm 1.31}$**	**78.34$_{\pm 1.71}$**
Multi-LLMs	Roberta	92.13$_{\pm 2.80}$	60.42$_{\pm 10.23}$	87.84$_{\pm 4.06}$	45.90$_{\pm 15.30}$	82.07$_{\pm 3.28}$	31.16$_{\pm 5.57}$	82.99$_{\pm 5.42}$	25.64$_{\pm 4.52}$	86.26$_{\pm 3.89}$	40.78$_{\pm 8.91}$
	LRR	92.78$_{\pm 0.27}$	26.20$_{\pm 0.00}$	94.98$_{\pm 0.27}$	86.20$_{\pm 0.00}$	81.94$_{\pm 0.05}$	39.80$_{\pm 0.00}$	59.93$_{\pm 0.41}$	0.10$_{\pm 0.00}$	82.41$_{\pm 0.25}$	38.07$_{\pm 0.00}$
	DetectGPT	54.92$_{\pm 0.13}$	0.11$_{\pm 0.14}$	59.05$_{\pm 0.22}$	0.02$_{\pm 0.06}$	55.29$_{\pm 0.37}$	0.08$_{\pm 0.06}$	55.98$_{\pm 0.35}$	0.00$_{\pm 0.00}$	56.31$_{\pm 0.27}$	0.05$_{\pm 0.06}$
	Fast-Detect.	94.98$_{\pm 1.72}$	71.90$_{\pm 0.00}$	95.97$_{\pm 2.65}$	94.20$_{\pm 0.00}$	93.19$_{\pm 1.48}$	81.80$_{\pm 0.00}$	56.57$_{\pm 3.33}$	0.40$_{\pm 0.00}$	85.18$_{\pm 2.30}$	62.08$_{\pm 0.00}$
	Str-Detect.	52.05$_{\pm 0.00}$	0.01$_{\pm 0.00}$	52.30$_{\pm 0.00}$	0.01$_{\pm 0.00}$	56.05$_{\pm 0.00}$	0.01$_{\pm 0.00}$	57.50$_{\pm 0.00}$	0.01$_{\pm 0.00}$	54.47$_{\pm 0.00}$	0.01$_{\pm 0.00}$
	Binoculars	97.95$_{\pm 0.89}$	85.90$_{\pm 0.00}$	98.05$_{\pm 1.17}$	98.70$_{\pm 0.00}$	**97.58$_{\pm 0.70}$**	**89.80$_{\pm 0.00}$**	68.59$_{\pm 3.55}$	3.40$_{\pm 0.00}$	90.54$_{\pm 1.58}$	69.45$_{\pm 0.00}$
	RepreGuard	**98.00$_{\pm 0.43}$**	**99.26$_{\pm 0.11}$**	**98.44$_{\pm 0.24}$**	**99.20$_{\pm 0.12}$**	92.35$_{\pm 0.46}$	74.74$_{\pm 4.52}$	**93.40$_{\pm 0.74}$**	**56.52$_{\pm 1.04}$**	**95.55$_{\pm 0.47}$**	**82.43$_{\pm 1.45}$**

Few-Shot Learning Demonstrates the Highest Overall Effectiveness

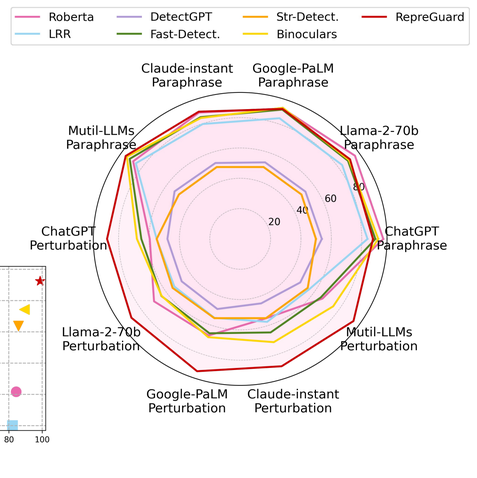
Shots→ / Train↓	Detector↓ / Metrics→	16 *AUR.*	16 *TPR.*	32 *AUR.*	32 *TPR.*	64 *AUR.*	64 *TPR.*	128 *AUR.*	128 *TPR.*	256 *AUR.*	256 *TPR.*	512 *AUR.*	512 *TPR.*	1024 *AUR.*	1024 *TPR.*
ChatGPT	Roberta	60.55	0.40	42.05	0.00	64.10	0.00	68.05	0.90	84.30	5.90	86.40	48.00	88.55	44.70
	LRR	80.55	31.50	81.10	31.50	82.30	31.50	82.85	31.50	83.35	31.50	83.45	31.50	82.75	31.50
	DetectGPT	53.68	0.15	54.07	0.04	54.02	0.19	55.89	0.08	56.22	0.13	57.88	0.06	57.78	0.17
	Fast-Detect.	**84.80**	58.80	**84.85**	58.80	84.75	58.80	84.70	58.80	84.75	58.80	84.70	58.80	84.70	58.80
	Binoculars	84.70	72.60	84.70	72.60	88.00	72.60	88.00	72.60	90.25	72.60	88.45	72.60	90.15	72.60
	RepreGuard	83.20	**81.00**	82.55	**83.70**	**91.05**	**81.60**	**90.50**	**81.00**	**93.15**	**78.60**	**93.20**	**79.70**	**93.85**	**79.40**
Llama-2-70b	Roberta	52.55	0.20	52.15	0.20	41.65	0.00	60.45	0.00	83.85	27.20	85.85	30.30	**95.50**	73.40
	LRR	75.40	31.50	81.10	31.50	82.80	31.50	82.95	31.50	82.45	31.50	82.45	31.50	82.80	31.50
	DetectGPT	54.12	0.11	53.89	0.07	54.33	0.18	56.01	0.02	55.76	0.16	57.23	0.09	58.09	0.13
	Fast-Detect.	79.35	58.80	84.10	58.80	84.20	58.80	84.50	58.80	85.00	58.80	84.70	58.80	84.70	58.80
	Binoculars	83.00	72.60	83.00	72.60	86.10	72.60	86.10	72.60	87.25	72.60	87.25	72.60	87.25	72.60
	RepreGuard	**89.55**	**79.20**	**89.55**	**78.70**	**96.35**	**77.90**	**91.85**	**82.00**	**95.50**	**81.70**	**94.35**	**81.80**	94.05	**82.70**
Google-PaLM	Roberta	55.45	0.50	61.10	0.10	66.30	0.00	71.25	2.40	77.50	2.80	89.40	22.90	**95.45**	64.50
	LRR	78.95	31.50	81.15	31.50	81.35	31.50	82.15	31.50	82.30	31.50	83.35	31.50	83.20	31.50
	DetectGPT	54.12	0.08	54.45	0.19	54.89	0.12	55.76	0.04	56.34	0.17	57.21	0.06	57.95	0.15
	Fast-Detect.	76.20	58.80	83.50	58.80	85.35	58.80	85.35	58.80	85.25	58.80	85.25	58.80	85.30	58.80
	Binoculars	86.00	72.60	88.95	72.60	88.95	72.60	88.95	72.60	88.95	72.60	88.45	72.60	89.85	72.60
	RepreGuard	**91.70**	**76.30**	**92.00**	**80.40**	**94.25**	**85.30**	**94.95**	**85.10**	**95.05**	**85.50**	**95.05**	**84.80**	95.30	**82.60**
Claude-instant	Roberta	51.90	0.00	62.30	0.00	50.85	0.00	62.10	0.20	74.45	9.30	84.85	43.90	83.70	66.00
	LRR	82.50	31.50	82.15	31.50	83.50	31.50	72.75	31.50	79.95	31.50	79.00	31.50	79.05	31.50
	DetectGPT	53.71	0.14	54.15	0.06	54.09	0.19	55.92	0.03	56.30	0.17	57.95	0.09	57.82	0.12
	Fast-Detect.	79.45	58.80	79.45	58.80	80.80	58.80	80.55	58.80	78.85	58.80	80.45	58.80	79.60	58.80
	Binoculars	82.15	**72.60**	82.15	**72.60**	82.15	**72.60**	82.15	72.60	82.15	72.60	81.90	72.60	85.80	72.60
	RepreGuard	**92.40**	63.20	**93.25**	72.50	**94.55**	69.90	**94.60**	**76.30**	**95.40**	**77.00**	**94.80**	**77.10**	**94.75**	**77.30**
Multi-LLMs	Roberta	57.25	0.00	50.45	1.60	59.80	0.70	59.60	1.30	64.95	3.60	93.50	34.80	95.65	46.90
	LRR	78.25	31.50	78.25	31.50	83.30	31.50	81.90	31.50	82.35	31.50	83.50	31.50	83.50	31.50
	DetectGPT	55.48	0.04	54.56	0.02	54.43	0.15	54.97	0.18	55.93	0.01	56.31	0.03	57.11	0.07
	Fast-Detect.	84.80	58.80	84.80	58.80	85.35	58.80	85.20	58.80	85.35	58.80	85.30	58.80	85.30	58.80
	Binoculars	82.15	72.60	82.15	72.60	89.30	72.60	89.30	72.60	89.45	72.60	90.45	72.60	90.45	72.60
	RepreGuard	**94.20**	**87.10**	**95.85**	**86.90**	**95.70**	**81.90**	**94.20**	**79.90**	**95.80**	**79.80**	**96.30**	**81.20**	**96.00**	**80.20**

Reliability in the Wild

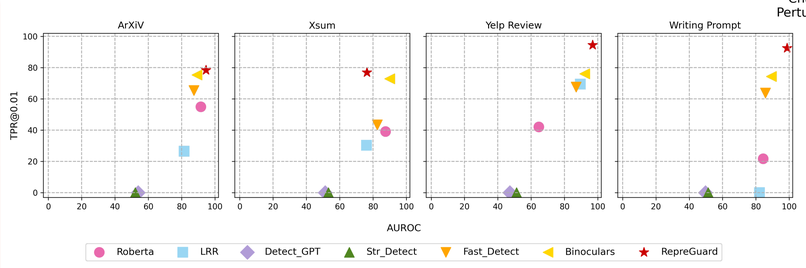
Efficient in Time and Space

Detector ↓	*AUR.*	*TPR.*	Cost of Space	Cost of Time (Per sample)
Roberta	84.85	43.90	**2.0GB**	**0.016s**
Fast-Detect.	80.45	58.80	40.0GB	0.390s
Binocular	81.90	72.60	58.0GB	0.653s
RepreGuard (Phi-2)	**96.10**	54.50	16.0GB	0.072s
RepreGuard (Llama-3.1-8B)	94.80	**77.10**	38.0GB	0.359s

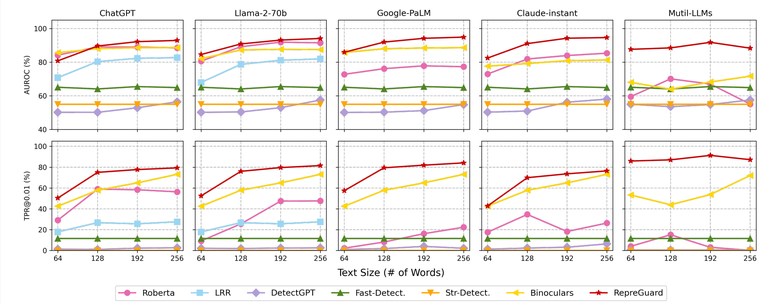
Resilience to Attacks



Robust Cross-Domain Generalization



Robustness to Varying Text Lengths



Robustness across Sampling Methods

	Chat Models (llama-chat, mistral-chat, mpt-chat)								Non-Chat Models (mistral, mpt, gpt2)								Avg.	
Dec. Strategy	greedy				sampling				greedy				sampling					
Rep. Penalty?																		
Metrics	*AUR.*	*TPR.*	*AUR.*	*TPR.*	*AUR.*	*TPR.*	*AUR.*	*TPR.*	*AUR.*	*TPR.*	*AUR.*	*TPR.*	*AUR.*	*TPR.*	*AUR.*	*TPR.*	*AUR.*	*TPR.*
Roberta	88.32	65.97	83.58	41.02	88.57	46.81	71.66	13.27	93.11	40.22	74.45	14.87	80.24	34.03	77.45	5.69	82.17	32.74
LRR	90.17	49.00	80.23	12.69	86.13	24.25	67.17	4.19	94.86	87.43	83.43	34.33	77.25	0.40	50.00	0.20	78.66	26.56
Fast-Detect.	97.80	95.21	87.56	68.03	97.06	91.12	71.86	30.04	98.65	96.21	77.50	48.10	84.03	31.44	50.00	0.00	83.06	57.52
Str-Detect.	55.99	0.01	55.14	0.01	55.24	0.01	55.34	0.01	56.94	0.01	53.84	0.01	56.64	0.01	55.69	0.01	56.98	0.01
Binoculars	**99.50**	**98.70**	91.52	71.26	**99.15**	94.41	77.69	31.24	**99.50**	**99.30**	79.54	33.43	**88.12**	1.70	50.05	0.00	85.63	53.76
RepreGuard	98.30	96.61	**97.16**	**94.81**	97.55	**94.61**	**94.86**	**85.73**	98.50	92.22	**92.47**	**75.55**	72.55	**34.43**	**81.99**	**46.31**	**92.05**	**77.53**

Data & Code can be found at: https://github.com/NLP2CT/RepreGuard

Our related works:

- Benchmarking LLM-Generated Text Detection in Real world Scenarios (NeurIPS 2024)
- A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions (CL 2025)
- Who Wrote This? The Key to Zero-Shot LLM-Generated Text Detection ls GECScore (COLING 2025)

[1]We use the DetectRL benchmark (Wu, et al., 2025), covering four domains: academic writing, news, creative writing, and social media. Here, TPR@0.01 denotes the true positive rate under an extremely low false positive rate (FPR = 0.01%)