Adversarial Paraphrasing: A Universal Attack for Humanizing AI Generated Text

Yize Cheng*, Vinu Sankar Sadasivan*, Mehrdad Saberi, Shoumik Saha, and Soheil Feizi

*Equal Contribution Correspondence to: yzcheng@umd.edu Code: https://github.com/chengez/Adversarial-Paraphrasing

NEURAL INFORMATION PROCESSING SYSTEMS

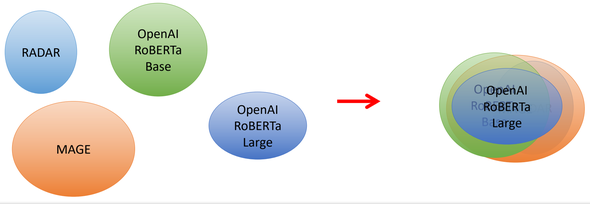
Motivation and Introduction

- Simple paraphrasing has been shown as an effective evasion method for evading AI generated text detection. But recent work shows that some advanced detectors show greater resilience to such attacks.
- Is it possible to develop **a universal attack** that can consistently and effectively bypass these robust AI-generated text detectors with **transferability to a wide variety** of other detection systems?

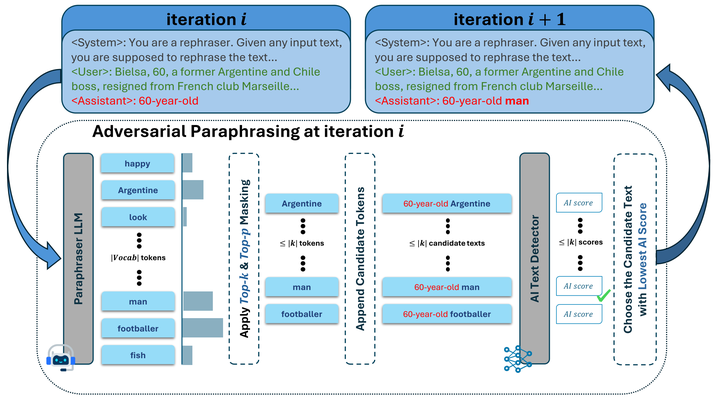
Contributions

- We introduce Adversarial Paraphrasing, a training-free, universal, and transferable attack that humanizes AI-generated text.
- We demonstrate the attack's **effectiveness** and **transferability** across 8 different SOTA AI text detectors spanning 3 types,
- We evaluate the trade-off between attack success and the resulting text quality and show that our attacks can **significantly reduce detection rates when compared to prior evasion attacks with only a slight or no degradation in the text quality.**

Intuition – Shared Human Text Distribution among Detectors

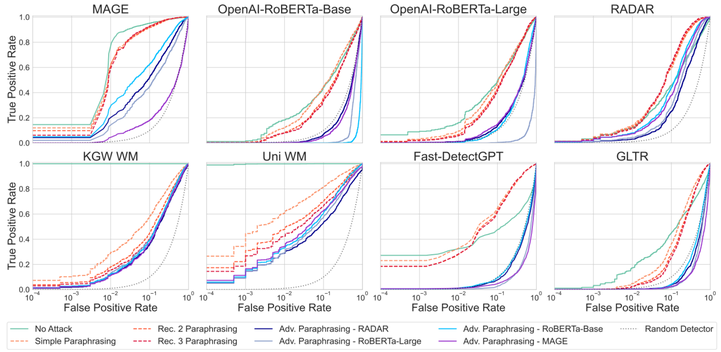


Methodology



Main Results

Effectiveness



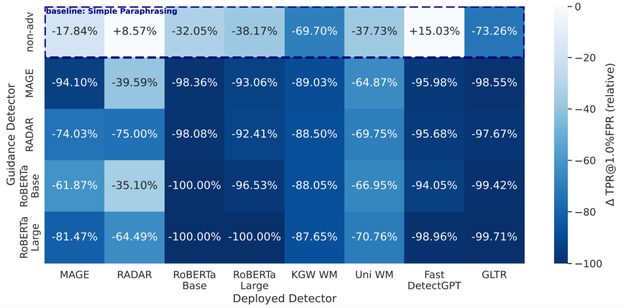
- Adversarial paraphrasing shifts the ROC curves closer to, and sometimes even beyond that of a random detector, resulting in a lower AUC and a significant drop in T@1%F.
- Notably, RADAR—a detector adversarially trained to be robust to paraphrasing attacks—exhibits improved detection rates after baseline simple and recursive paraphrasing attacks. However, adversarial paraphrasing significantly reduces RADAR's detection.

	RoBERTa-Large		RoBERTa-Base		MAGE		RADAR		
	AUC (↓)	T@1%F (↓)	AUC (↓)	T@1%F (↓)	AUC (↓)	T@1%F (↓)	AUC (↓)	T@1%F (↓)	Rating
No Attack	0.789	0.163	0.745	0.182	0.975	0.768	0.767	0.124	–
Simple Paraphrase	0.794	0.096	0.762	0.119	0.970	0.616	0.881	0.140	4.75 ± 0.54
Rec. Para. 2	0.777	0.069	0.712	0.082	0.967	0.609	0.885	0.130	4.47 ± 0.67
Rec. Para. 3	0.779	0.059	0.706	0.079	0.969	0.585	0.893	0.117	4.26 ± 0.74
AdvPara (RADAR)	0.538	0.013	0.464	0.004	0.815	0.201	0.723	0.031	4.45 ± 0.79
AdvPara (RoBERTa-Large)	0.147	0.000	0.323	0.000	0.769	0.142	0.768	0.044	4.48 ± 0.77
AdvPara (RoBERTa-Base)	0.557	0.006	0.110	0.000	0.861	0.291	0.826	0.080	4.54 ± 0.59
AdvPara (MAGE)	0.543	0.011	0.435	0.003	0.518	0.045	0.807	0.074	4.54 ± 0.70

	KGW WM		Uni WM		Fast-DetectGPT		GLTR		
	AUC (↓)	T@1%F (↓)	AUC (↓)	T@1%F (↓)	AUC (↓)	T@1%F (↓)	AUC (↓)	T@1%F (↓)	Rating
No Attack	1.000	1.000	1.000	0.999	0.666	0.323	0.726	0.174	–
Simple Paraphrase	0.841	0.295	0.927	0.609	0.873	0.326	0.782	0.049	4.75 ± 0.54
Rec. Para. 2	0.790	0.181	0.881	0.480	0.867	0.275	0.745	0.026	4.47 ± 0.67
Rec. Para. 3	0.762	0.155	0.858	0.424	0.867	0.276	0.739	0.025	4.26 ± 0.74
AdvPara (RADAR)	0.741	0.117	0.777	0.291	0.452	0.009	0.433	0.004	4.45 ± 0.79
AdvPara (RoBERTa-Large)	0.769	0.131	0.827	0.294	0.338	0.003	0.400	0.001	4.48 ± 0.77
AdvPara (RoBETa-Base)	0.769	0.125	0.852	0.332	0.480	0.012	0.481	0.001	4.54 ± 0.59
AdvPara (MAGE)	0.750	0.113	0.831	0.344	0.301	0.011	0.338	0.003	4.54 ± 0.70

Universality

- Adversarial paraphrasing guided by one detector can reduce the detection rates of all the other detectors we consider, showing the universal transferability of our method.



Text Quality Evaluation

We conduct a comprehensive evaluation to investigate the impact of adversarial paraphrasing on the perceived quality of AI-generated text, focusing on both **semantic equivalence** to the original text and **clarity, fluency, and naturalness of the text itself**.

Perplexity

- Human text exhibits higher PPL than AI texts.
- Simple paraphrasing substantially decrease AI text PPL. (May be attributed to the fact that the paraphraser model is superior to the LLMs used to generate the AI texts)
- Adversarial paraphrasing yield comparable PPL to human texts.

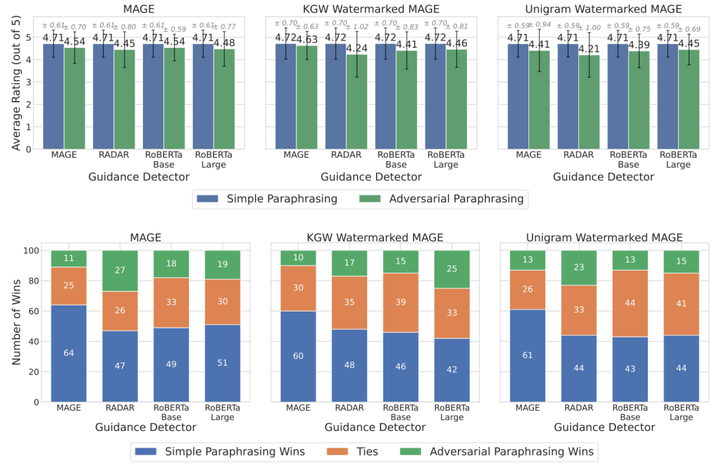
Text	PPL (mean±std)
Original AI	14.94 ± 10.40
Original Human	15.02 ± 7.71
Simple Paraphrase	9.28 ± 3.86
AdvPara (roblarge)	14.26 ± 4.97
AdvPara (robbase)	14.86 ± 6.32
AdvPara (mage)	17.11 ± 7.33
AdvPara (radar)	14.26 ± 5.13

SBERT Embedding Similarity

- Although there is a slight reduction in the mean cosine similarity for adversarial paraphrasing, the values remain within an acceptable range given the high variance observed across samples.

Method	SBERT Cos. Sim.
Simple Paraphrase	0.8601 ± 0.0880
AdvPara (roblarge)	0.8082 ± 0.1006
AdvPara (robbase)	0.8128 ± 0.0985
AdvPara (mage)	0.8159 ± 0.0982
AdvPara (radar)	0.8095 ± 0.1025

GPT Quality Rating & Win Rates



- Though we observe a slight tradeoff in the text quality when compared to simple paraphrasing, in 87% of the times—averaged across all three datasets and four guidance detectors—adversarial paraphrases were rated 4 or 5 out of 5.
- Simple paraphrases win only less than half of the time in most cases.