Stumbling Blocks: Stress Testing the Robustness of Machine-Generated Text Detectors Under Attacks

Yichen Wang[2,*], Shangbin Feng[1], Abe Bohan Hou[3], Xiao Pu[4],
Chao Shen[2], Xiaoming Liu[2], Yulia Tsvetkov[1], Tianxing He[1]

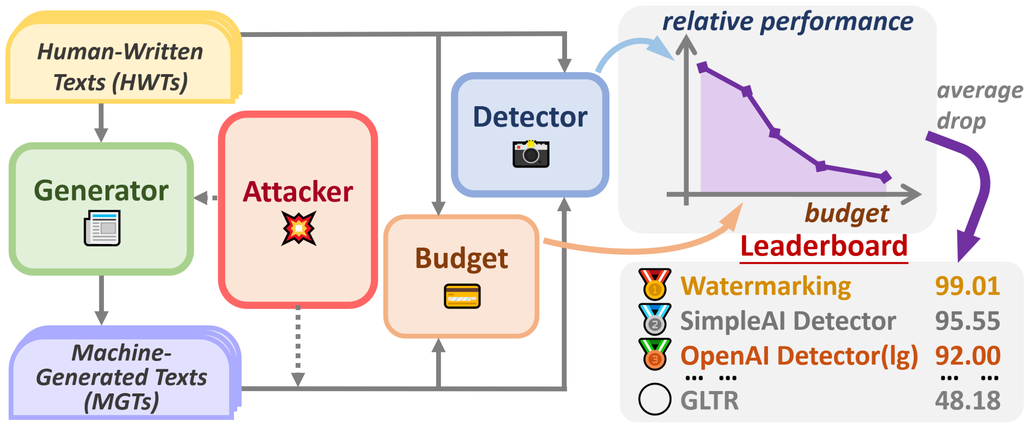
* Done while interning at UW.

[1] Paul G. Allen School of CSE, University of Washington
[2] Xi'an Jiaotong University [3]Johns Hopkins University [4]Peking University

Comprehensively **benchmark and study** the robustness of 8 prevalent machine-generated text detectors under 12 malicious attacks.

🏆 leaderboard 🕵️ defect analysis 👨‍🏫 interpretation 🔮 defense patch

Scope: - Attacker does not have any knowledge/access to the detectors;
- Attacker only has limited access to the generators (OAI panel-like);
- Apply each attack on different perturbation levels, termed as budgets.



Detectors: Fine-Tuned Detector (OAI detector, tuned DeBERTa ...)
Watermark-Based Detector (Kirchenbauer et al. 2023a)
Metric-Based Detector (GLTR, Rank, DetectGPT ...)

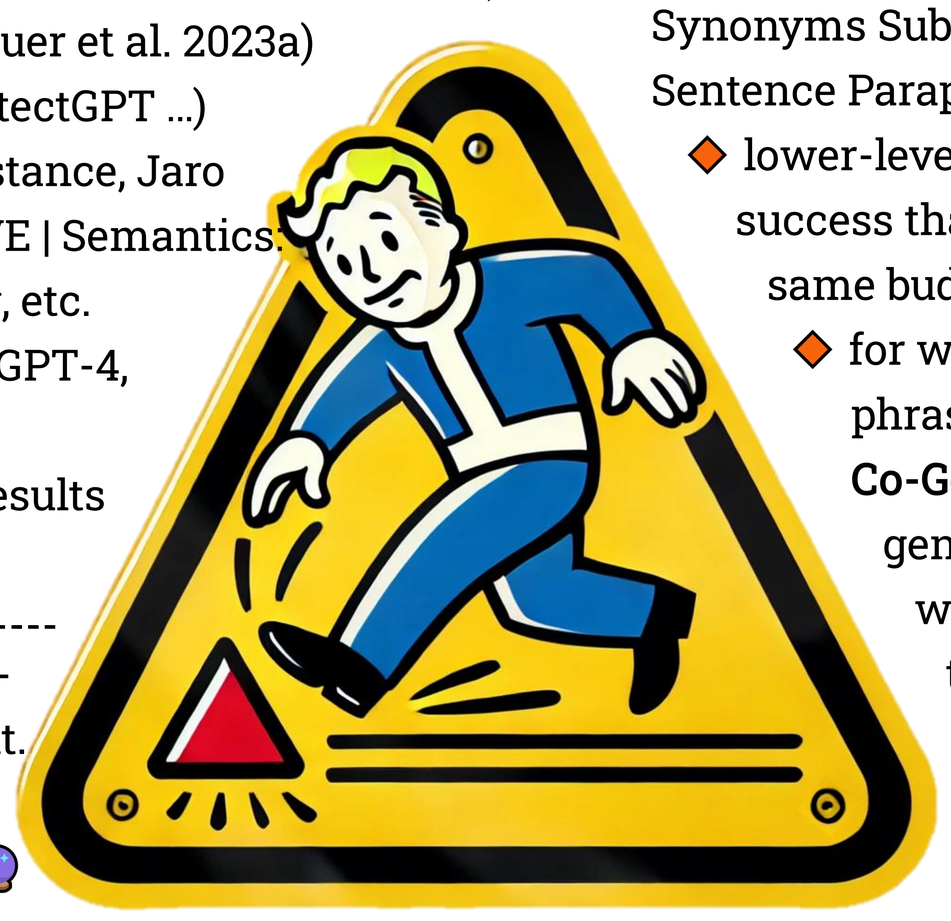
Budgets: Editing - Levenshtein Edit Distance, Jaro Similarity | Quality – Perplexity, MAUVE | Semantics: BERTScore, BARTScore, Cos. Similarity, etc.

Generators: GPT-J-6B, LlaMA-2-7B-hf, GPT-4, Davinci-003, LlaMA-1, etc.

✨ All the generators shared **similar** results under attacks.

Editing Attacks: Typo Insertion, Homoglyph Alteration, Format Character Edit.

◆ metric-based methods perform the worst. Most f.t. detectors fail. 👨‍🏫 🔮



Paraphrasing Attacks: cover word- to paragraph-level Synonyms Substitution, Span Perturbation, Inner-Sentence Paraphrase, Inter-Sentence Paraphrase.

◆ lower-level perturbations show greater attack success than higher-level perturbations at the same budget.

◆ for watermarking, inter-sentence paraphrasing is the only effective attack. 👨‍🏫 🔮

Co-Generating Attacks: perturbs the generated tokens at each recurrent step with some designed rules. E.g., co-gen. typo💥, emoji. 👨‍🏫

Prompting Attacks: Prompt Paraphrasing, ICL, Character-Substituted Generation💥. 👨‍🏫 🔮

Takeaways 🤔 Can the current MGT detector robustly detect?

- **Almost none of the existing detectors remains robust under all attacks.** Averaging all detectors, the performance drops by 35% across all attacks.
  - E.g., about 2 to 6 character editing by typo insertion can severely deceive metric-based detectors (e.g., DetectGPT), to perform worse than a random prediction. (average length is around 120 tokens)
- **Watermarking performs best** to its applicable attacks, but still fails on *inter-sentence paraphrasing* attacks, etc.
- **Model-based detectors** are **more robust** than metric-based ones in most cases. (Among which SimpleAI det. is best.)

Experiment Results: Leaderboard

Leaderboard: MGT Detector Robustness

Detector	*Edit*	*Para.*	*Prompt*	*CoGen.*	*Avg.*
Watermark	**99.86**	**97.17**	- -	**99.99**	**99.01***
SimpleAI Det.	**108.1**	97.51	81.58	95.04	**95.55**
OpenAI Det.-Lg	57.77	**97.84**	**105.2**	**107.2**	92.00
Model. Avg.	76.65	92.08	97.57	92.22	89.63
F.t. DeBERTa	104.1	81.49	99.09	64.28	87.24
OpenAI Det.-Bs	36.63	91.46	104.4	102.4	83.71
DetectGPT-1d	**74.82**	**75.32**	**102.8**	**66.46**	**79.85**
DetectGPT-10d	62.67	64.40	97.68	49.78	68.63
DetectGPT-10z	56.41	59.73	93.88	43.08	63.28
Metric. Avg.	51.82	61.89	91.26	33.49	59.62
LogRank	41.76	58.38	84.44	11.20	48.95
Rank	36.46	57.68	81.00	20.08	48.81
GLTR	38.82	55.80	87.79	10.32	48.18

Experiment Results: Performance Degradation

More Results, Details and Info:

Paper

Bench mark

ACL 2024 Bangkok, Thailand

Personal News: I'll be a Ph.D. student at UChicago this fall, to be advised by Prof. *Mina Lee* and Prof. *Ari Holtzman*. New papers are coming up! ↩