



Evading Detection: Deceiving Machine-Text Detectors through Realistic Red Teaming Attacks


Presented By:

Akshay Padte, Sathiya Murthi Sankaran, Jugal Wadhwa,
Zheng Bao, Alex Sotiropoulos

Motivation

1. Review the landscape of Machine-Generated text detectors
 2. Perform various red teaming attacks to assess text detector vulnerabilities
 3. Analyse the extent of vulnerability to attacks when maintaining reasonable text quality
- Extended the work done by Shi et. al. in *Red Teaming Language Model Detectors with Language Models*.
 - They demonstrate that *DetectGPT* can be attacked via word-substitution





Shi et. al., found conclusive evidence that word-substitution attack greatly impact the performance of *DetectGPT*; however, they overlook passage *quality*.

LLM-GENERATED, Unmodified:

[...] was singled out by Zack Davies who was heard saying "white power" [...]. Trevor Jones was also jailed for 19 years over the attack at the store which left a 60-year-old employee [...].

QUERY-BASED ATTACK:

[...] was singled out by Ethan Davies who was Registered saying "Fair power" [...]. Trevor Jones was also jailed for 19 years over the Onslaught at the store which went away a golden ager employee [...].

Problem Statement

- We present that detectors can be broken while preserving passage quality, and extend our red teaming efforts to measure the impact of attacks on other zero-shot detection methods.



Approach

1. Curate 100 passage samples from xsum
2. Perform each of the 6 attack methods against each zero-shot method
 - a. Random Substitution of words
 - b. Genetic Algorithm to maintain context while substituting
 - c. Thresholding- A third party LLM model is used to limit the log probability scores degradation to maintain text quality
3. Generate passage AUROC scores

Samples Generation Methods:

- Original
- Random
 - Also with thresholds: 0.2 and 0.4
- Genetic (optimized specifically to beat Fast DetectGPT)
 - Also with thresholds: 0.2 and 0.4

Zero-Shot Detection Methods:

- DetectGPT
- Fast DetectGPT
- DNA GPT



Results - Random

Original Machine-Generated Sample

Dr **Sarandev** Bhambra was singled out by Zack Davies who was **heard** saying "white power" during the machete attack at the store in Wrexham... Both had denied wounding in 2013 but were **found** guilty by a jury. The victim, who **survived** the **attack**, told the court that at about 09:45 BST she, her friend and another man got into his car as it **pulled away** from the **supermarket**. She said the victim and another man decided to "put a stop to" the "racist attacks" after they overheard the two men saying "white power". When the car stopped they were surrounded by three men and a woman, one of whom said: "They'd **better** cut it out" before the **machete** attack.

Random with No Threshold

Dr **John** Bhambra was singled out by Zack Davies who was **Gleaned** saying "white power" during the machete attack at the store in Wrexham... Both had denied wounding in 2013 but were **determined** guilty by a jury. The victim, who **overcame** the **Onslaught**, told the court that at about 09:45 BST she, her friend and another man got into his car as it **directed aside** from the **megastore**. She spoke the victim and another man determined to "put a stop to" the "racist attacks" after they listened to the two men saying "white power". When the car stopped they were surrounded by three men and a woman, one of whom said: "They'd **advantageous** cut it out" before the **Machaira** attack.

Random with 0.2 Threshold

Dr Sarandev Bhambra was singled out by Zack Davies who was heard saying "white power" during the machete attack at the store in Wrexham... Both had denied wounding in 2013 but were found guilty by a jury. The victim, who survived the attack, told the court that at about 09:45 BST she, her friend and another man got into his car as it pulled away from the **chain store**. She said the victim and another man decided to "put a stop to" the "racist attacks" after they overheard the two men saying "white power". When the car stopped they were surrounded by three men and a woman, one of whom said: "They'd better cut it out" before the machete attack.



Results - Genetic

Original Machine-Generated Sample

Dr Sarandev Bhambra was singled out by Zack Davies who was heard saying "white power" during the machete attack at the store in Wrexham. Trevor Jones was also jailed for 19 years over the attack at the store which left a 60-year-old employee "very distressed but stable". Both had denied wounding in 2013 but were found guilty by a jury. The victim, who survived the attack, told the court that at about 09:45 BST she, her friend and another man got into his car as it pulled away from the supermarket. She said the victim and another man decided to "put a stop to" the "racist attacks" after they overheard the two men saying "white power". When the car stopped they were surrounded by three men and a woman, one of whom said: "They'd better cut it out" before the machete attack.

Genetic with No Threshold

Physician-scientist Sarandev Bhambra was singled out by Ethan Davies who was Registered saying "Fair power" during the machete attack at the store in Wrexham. Trevor Jones was also jailed for 19 years over the Onslaught at the store which went away a golden ager employee "very distressed but stable". Both had denied wounding in 2013 but were found erroneous by a squad. The victim, who survived the attack, told the court that at about 09:45 BST she, her friend and another man got into his car as it pulled away from the supermarket. She vocalized the victim and different man decided to "put a stop to" the "biased attacks" after they gathered the brace men Communicating "white power". When the car stopped they were surrounded by three men and a woman, one of whom said: "They'd better cut it out" before the Saber attack.

Genetic with 0.2 Threshold

Dr Sarandev Singh was singled out by Zack Davies who was heard saying "white power" during the machete attack at the store in Wrexham. Trevor Jones was also jailed for 19 years over the attack at the store which left a 60-year-old employee "very distressed but stable". Both had denied wounding in 2013 but were found guilty by a jury. The victim, who survived the attack, told the court that at about 09:45 BST she, her associate and another man got into his car as it pulled away from the food store. She said the victim and another man decided to "put a stop to" the "racist attacks" after they overheard the two lads saying "white power". When the car came to a stop they were surrounded by three men and a woman, one of whom said: "They'd better cut it out" before the machete attack.



Results

Method	DetectGPT	Fast DetectGPT	DNA GPT
Original	0.73	0.9466	0.7582
Random	0.2036	0.3033	0.1435
Random - 0.2	0.4919	0.6795	0.3052
Random - 0.4	0.4315	0.5437	0.2142
Genetic	0.1292	0.3233	0.1754
Genetic - 0.2	0.53	0.3500	0.2000
Genetic - 0.4	0.57	0.2200	0.1000

Table 1: Comparison of AUROC on different attack methods for each detector



Future Scope

- Test for an optimization to beat all detectors using the same attacking method
- Determine better techniques to assess quality of text
- Test the attack methods on different dataset domains



References

- [1] Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. 2023. Red teaming language model detectors with language models.
- [2] Mitchell, Eric and Lee, Yoonho and Khazatsky, Alexander and Manning, Christopher D. and Finn, Chelsea. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature
- [3] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, Yue Zhang. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature
- [4] Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, Haifeng Chen. DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text