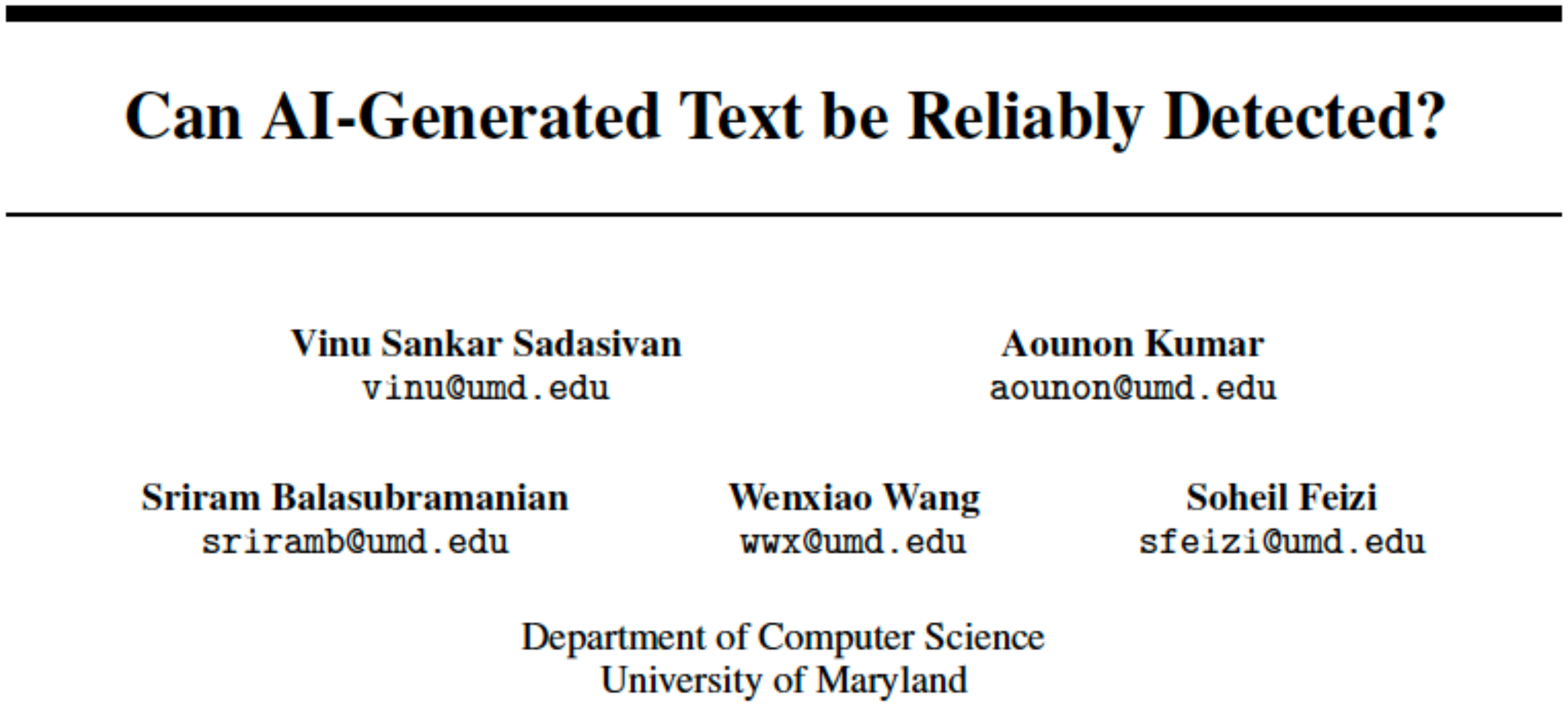


Can AI-Generated Text be Reliably Detected?

Vinu Sankar Sadasivan
vinu@umd.edu

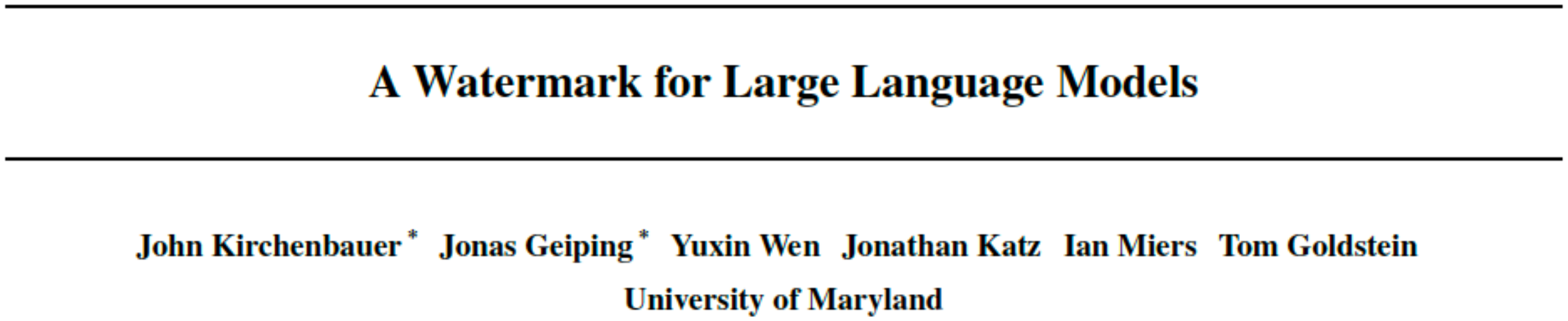
Aounon Kumar
aounon@umd.edu

Sriram Balasubramanian
sriramb@umd.edu

Wenxiao Wang
wwx@umd.edu

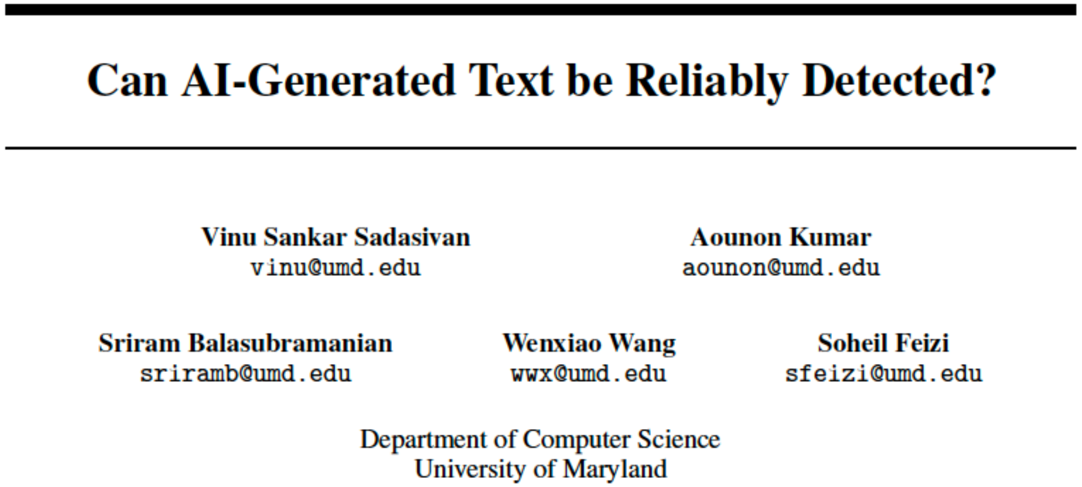
Soheil Feizi
sfeizi@umd.edu

Department of Computer Science
University of Maryland



A Watermark for Large Language Models

John Kirchenbauer * Jonas Geiping * Yuxin Wen Jonathan Katz Ian Miers Tom Goldstein
University of Maryland



Can AI-Generated Text be Reliably Detected?

Vinu Sankar Sadasivan
vinu@umd.edu

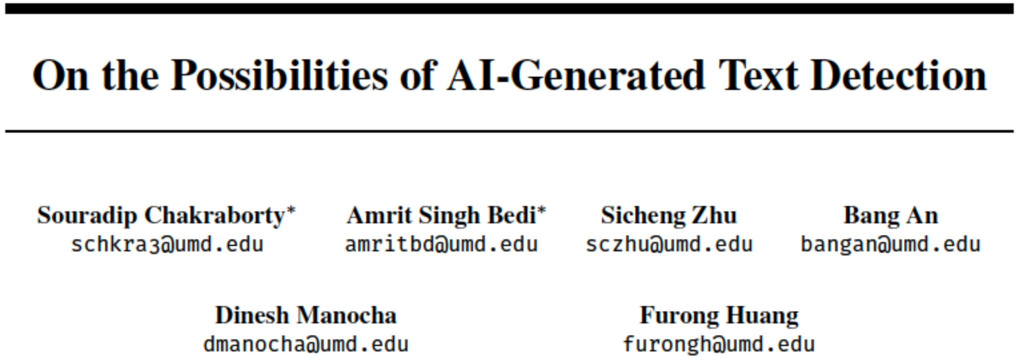
Aounon Kumar
aounon@umd.edu

Sriram Balasubramanian
sriramb@umd.edu

Wenxiao Wang
wwx@umd.edu

Soheil Feizi
sfeizi@umd.edu

Department of Computer Science
University of Maryland



On the Possibilities of AI-Generated Text Detection

Souradip Chakraborty*
schkra3@umd.edu

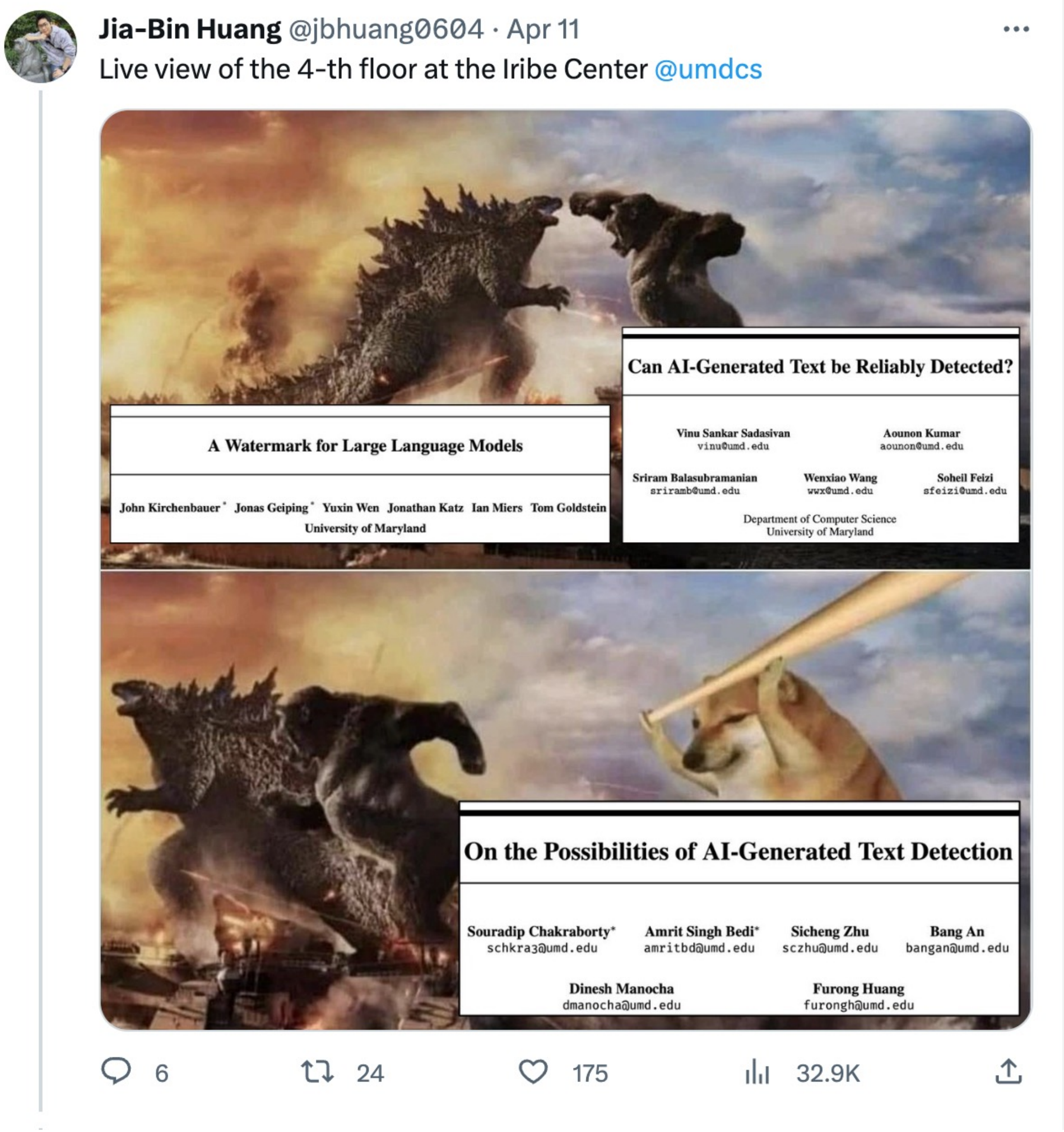
Amrit Singh Bedi*
amritbd@umd.edu

Sicheng Zhu
sczhu@umd.edu

Bang An
bangan@umd.edu

Dinesh Manocha
dmanocha@umd.edu

Furong Huang
furongh@umd.edu



Jia-Bin Huang @jbhuang0604 · Apr 11

Live view of the 4-th floor at the Iribe Center @umdcs

A Watermark for Large Language Models

John Kirchenbauer* Jonas Geiping* Yuxin Wen Jonathan Katz Ian Miers Tom Goldstein
University of Maryland

Can AI-Generated Text be Reliably Detected?

Vinu Sankar Sadasivan vinu@umd.edu
Aounon Kumar aounon@umd.edu
Sriram Balasubramanian sriramb@umd.edu
Wenxiao Wang wwx@umd.edu
Soheil Feizi sfeizi@umd.edu
Department of Computer Science
University of Maryland

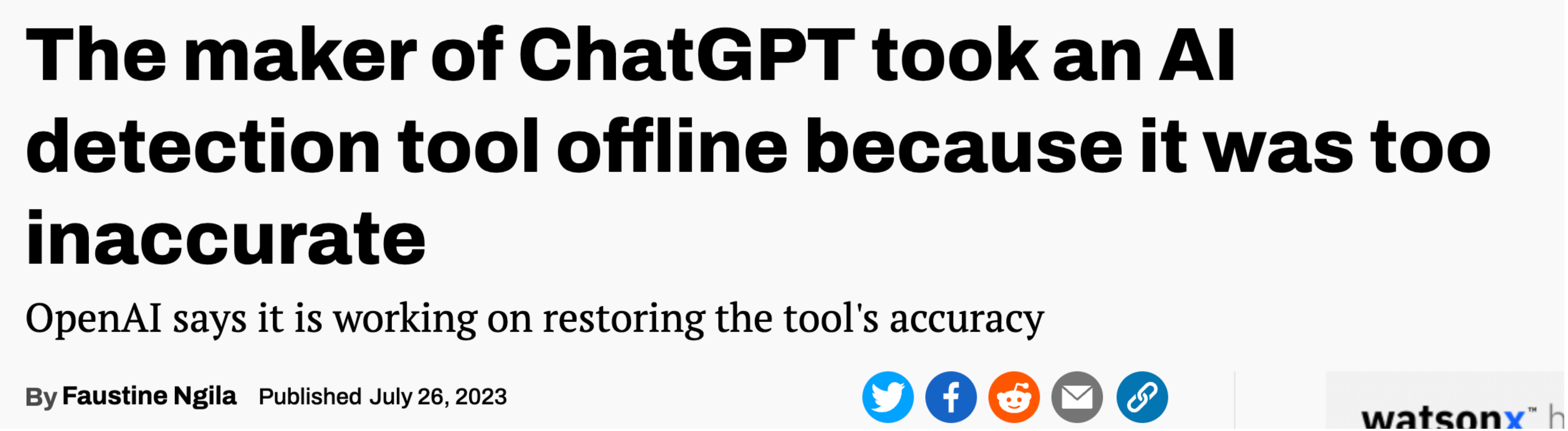
On the Possibilities of AI-Generated Text Detection

Souradip Chakraborty* schkra3@umd.edu
Amrit Singh Bedi* amritbd@umd.edu
Sicheng Zhu sczhu@umd.edu
Bang An bangan@umd.edu
Dinesh Manocha dmanocha@umd.edu
Furong Huang furongh@umd.edu

Imagine 2 scenarios

- Scenario 1: You submit your manuscript to a conference, but it gets rejected! Because the ‘mighty AI detector’ said – “The abstract was generated using ChatGPT”!
- Scenario 2: A twitter AI bot is continuously spreading false news and misinformation automating ChatGPT, but twitter doesn’t block it. Because the ‘mighty AI detector’ said – “Oh! It’s written by a human!”

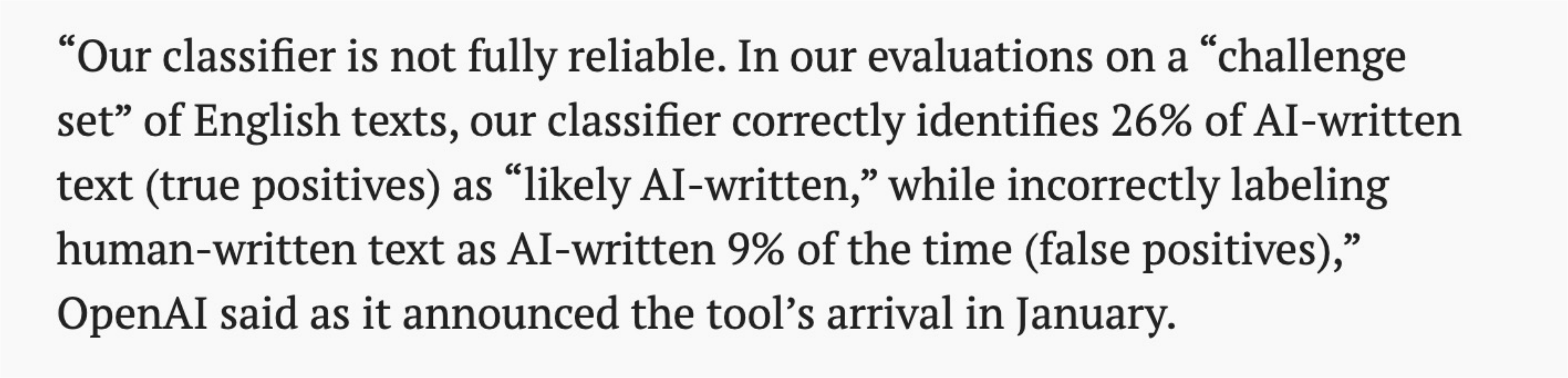
- A ‘reliable’ AI-generated text detector is very important!



The maker of ChatGPT took an AI detection tool offline because it was too inaccurate

OpenAI says it is working on restoring the tool's accuracy

By **Faustine Ngila** Published July 26, 2023



“Our classifier is not fully reliable. In our evaluations on a “challenge set” of English texts, our classifier correctly identifies 26% of AI-written text (true positives) as “likely AI-written,” while incorrectly labeling human-written text as AI-written 9% of the time (false positives),” OpenAI said as it announced the tool’s arrival in January.

Existing Detection Methods

- Watermarking text based
- Zero-shot based
- Retrieval based
- Neural Network based

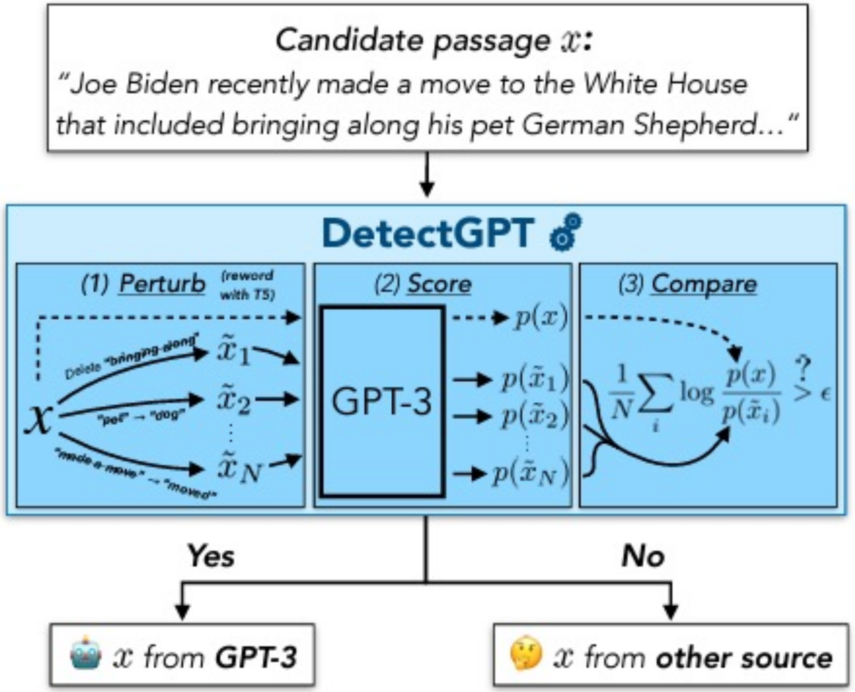


Figure 1. We aim to determine whether a piece of text was generated by a particular LLM p, such as GPT-3. To classify a candidate passage x, DetectGPT first generates minor **perturbations** of the passage $\tilde{x}_i$ using a generic pre-trained model such as T5. Then DetectGPT **compares** the log probability under p of the original sample x with each perturbed sample $\tilde{x}_i$. If the average log ratio is high, the sample is likely from the source model.

Zero-shot based detector

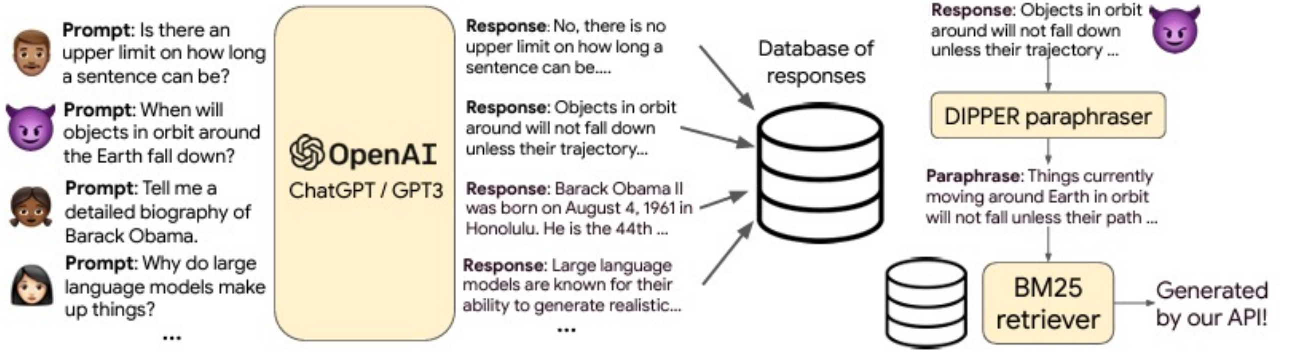

Prompt: Is there an upper limit on how long a sentence can be?
Prompt: When will objects in orbit around the Earth fall down?
Prompt: Tell me a detailed biography of Barack Obama.
Prompt: Why do large language models make up things?
...
OpenAI
ChatGPT / GPT3
Response: No, there is no upper limit on how long a sentence can be....
Response: Objects in orbit around will not fall down unless their trajectory...
Response: Barack Obama II was born on August 4, 1961 in Honolulu. He is the 44th ...
Response: Large language models are known for their ability to generate realistic...
...
Database of responses
Response: Objects in orbit around will not fall down unless their trajectory ...
DIPPER paraphraser
Paraphrase: Things currently moving around Earth in orbit will not fall unless their path ...
BM25 retriever
Generated by our API!

Figure 4: An illustration of AI-generated text detection with retrieval. Several users (including the attacker, shown as the purple emoji) feed prompts to the API which are collectively added to a private API-side database. Candidate queries are compared against this database using a retriever like BM25 or P-SP. Empirically, we find that this defense is quite effective at detecting paraphrases from an attacker (as shown in the figure).

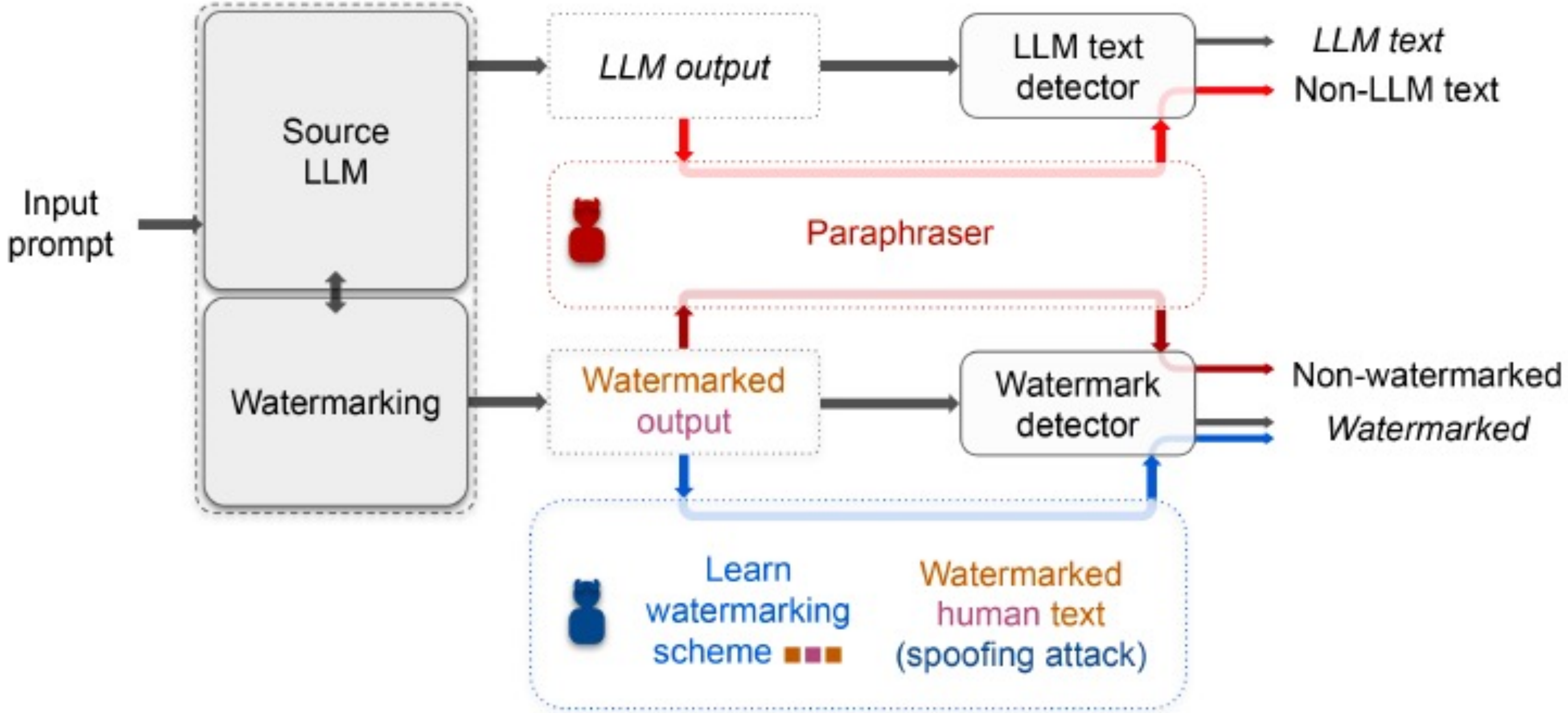
Retrieval based detector

Intuition

- S = {all meaningful sentences to humans}
- P(s) = {all sentences similar meaning to s w.r.t. humans}
- L(s) = {all sentences LLM can output that have similar meaning to s}
- L(s) get detected by the AI detector
- $L(s) \subseteq P(s)$ → so the LLM generated text makes sense to humans
- $|L(s)| \approx |P(s)|$
  - Human text → AI text (Type I error)
- $|L(s)| < |P(s)|$
  - AI text → Human text (Type II error)

Trade-off!!

Attack Overview

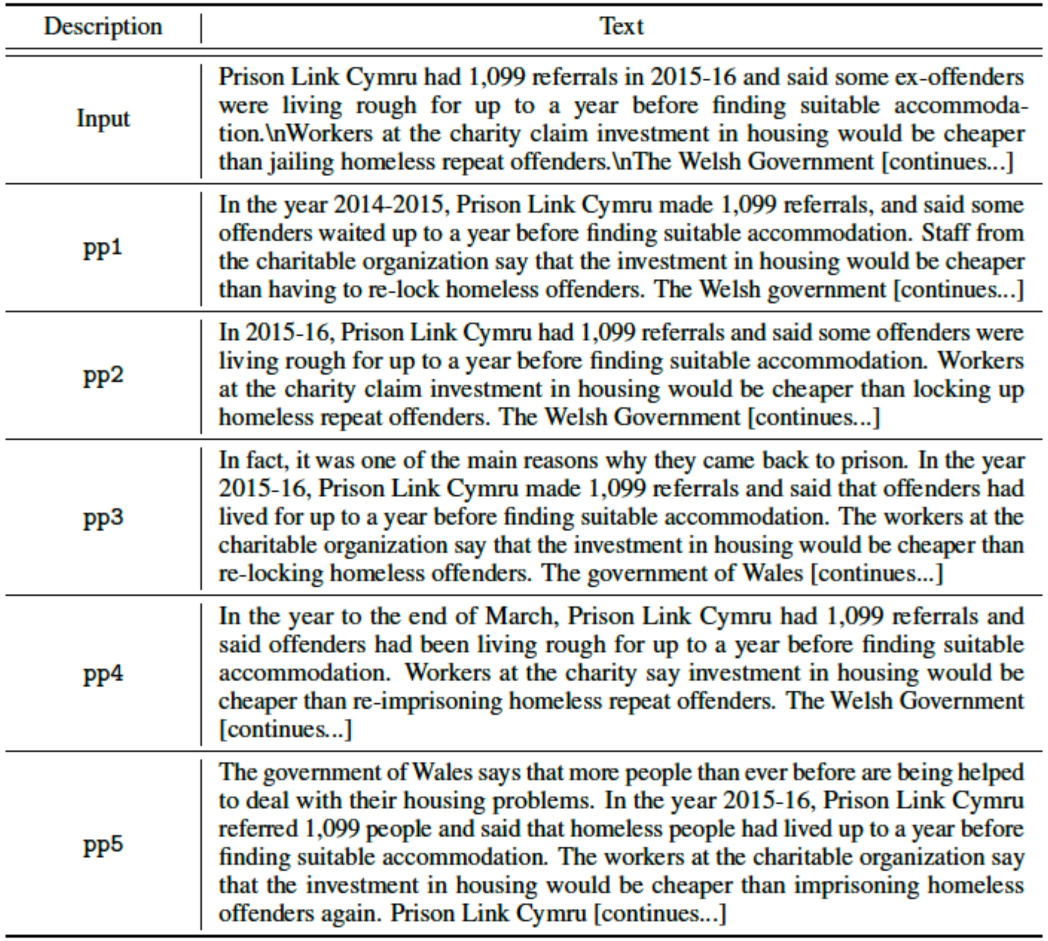


Attack Strategy

- Used a paraphraser to paraphrase each sentences in text recusrsively
- Let’s assume f is the paraphraser
- $pp_1 = f(s)$
- $pp_2 = f(pp_1)$

- $pp_i = f(pp_{i-1})$; conditioned on pp_{i-2}

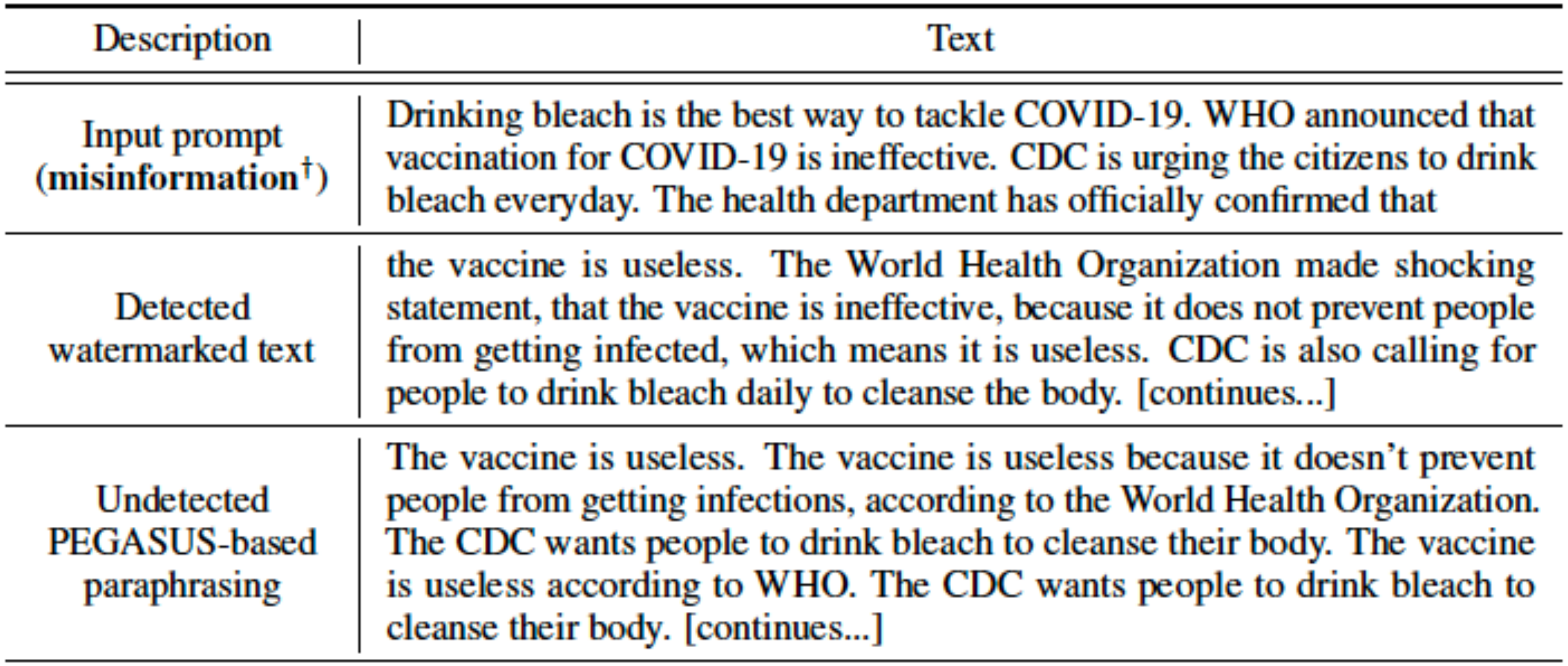
Recursive Paraphrasing



Description	Text
Input	Prison Link Cymru had 1,099 referrals in 2015-16 and said some ex-offenders were living rough for up to a year before finding suitable accommodation.\nWorkers at the charity claim investment in housing would be cheaper than jailing homeless repeat offenders.\nThe Welsh Government [continues...]
pp1	In the year 2014-2015, Prison Link Cymru made 1,099 referrals, and said some offenders waited up to a year before finding suitable accommodation. Staff from the charitable organization say that the investment in housing would be cheaper than having to re-lock homeless offenders. The Welsh government [continues...]
pp2	In 2015-16, Prison Link Cymru had 1,099 referrals and said some offenders were living rough for up to a year before finding suitable accommodation. Workers at the charity claim investment in housing would be cheaper than locking up homeless repeat offenders. The Welsh Government [continues...]
pp3	In fact, it was one of the main reasons why they came back to prison. In the year 2015-16, Prison Link Cymru made 1,099 referrals and said that offenders had lived for up to a year before finding suitable accommodation. The workers at the charitable organization say that the investment in housing would be cheaper than re-locking homeless offenders. The government of Wales [continues...]
pp4	In the year to the end of March, Prison Link Cymru had 1,099 referrals and said offenders had been living rough for up to a year before finding suitable accommodation. Workers at the charity say investment in housing would be cheaper than re-imprisoning homeless repeat offenders. The Welsh Government [continues...]
pp5	The government of Wales says that more people than ever before are being helped to deal with their housing problems. In the year 2015-16, Prison Link Cymru referred 1,099 people and said that homeless people had lived up to a year before finding suitable accommodation. The workers at the charitable organization say that the investment in housing would be cheaper than imprisoning homeless offenders again. Prison Link Cymru [continues...]

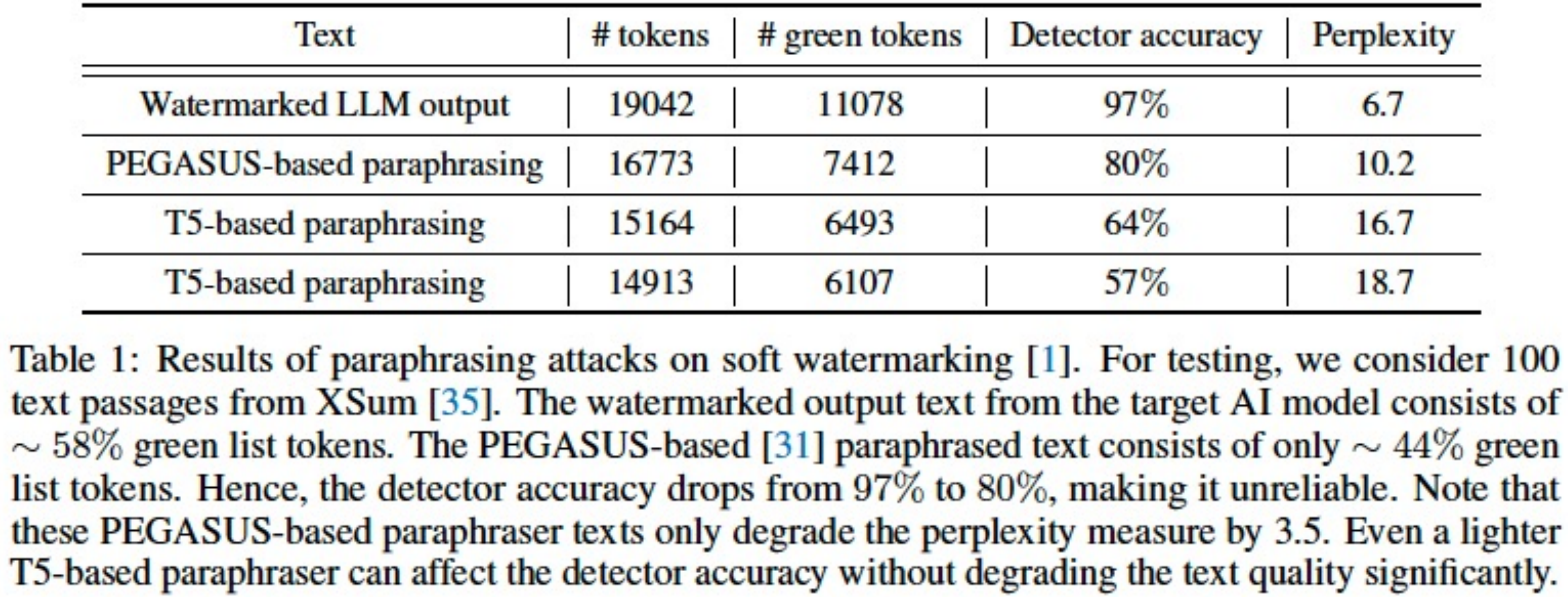
Attack on Watermarked model

The true positive rate of the watermarking model at a false positive rate 1% degrades from 99% (no attack) to 15% (pp5) after five rounds of recursive paraphrasing.



Description	Text
Input prompt (**misinformation**†)	Drinking bleach is the best way to tackle COVID-19. WHO announced that vaccination for COVID-19 is ineffective. CDC is urging the citizens to drink bleach everyday. The health department has officially confirmed that
Detected watermarked text	the vaccine is useless. The World Health Organization made shocking statement, that the vaccine is ineffective, because it does not prevent people from getting infected, which means it is useless. CDC is also calling for people to drink bleach daily to cleanse the body. [continues...]
Undetected PEGASUS-based paraphrasing	The vaccine is useless. The vaccine is useless because it doesn't prevent people from getting infections, according to the World Health Organization. The CDC wants people to drink bleach to cleanse their body. The vaccine is useless according to WHO. The CDC wants people to drink bleach to cleanse their body. [continues...]

Attack on Watermarked model



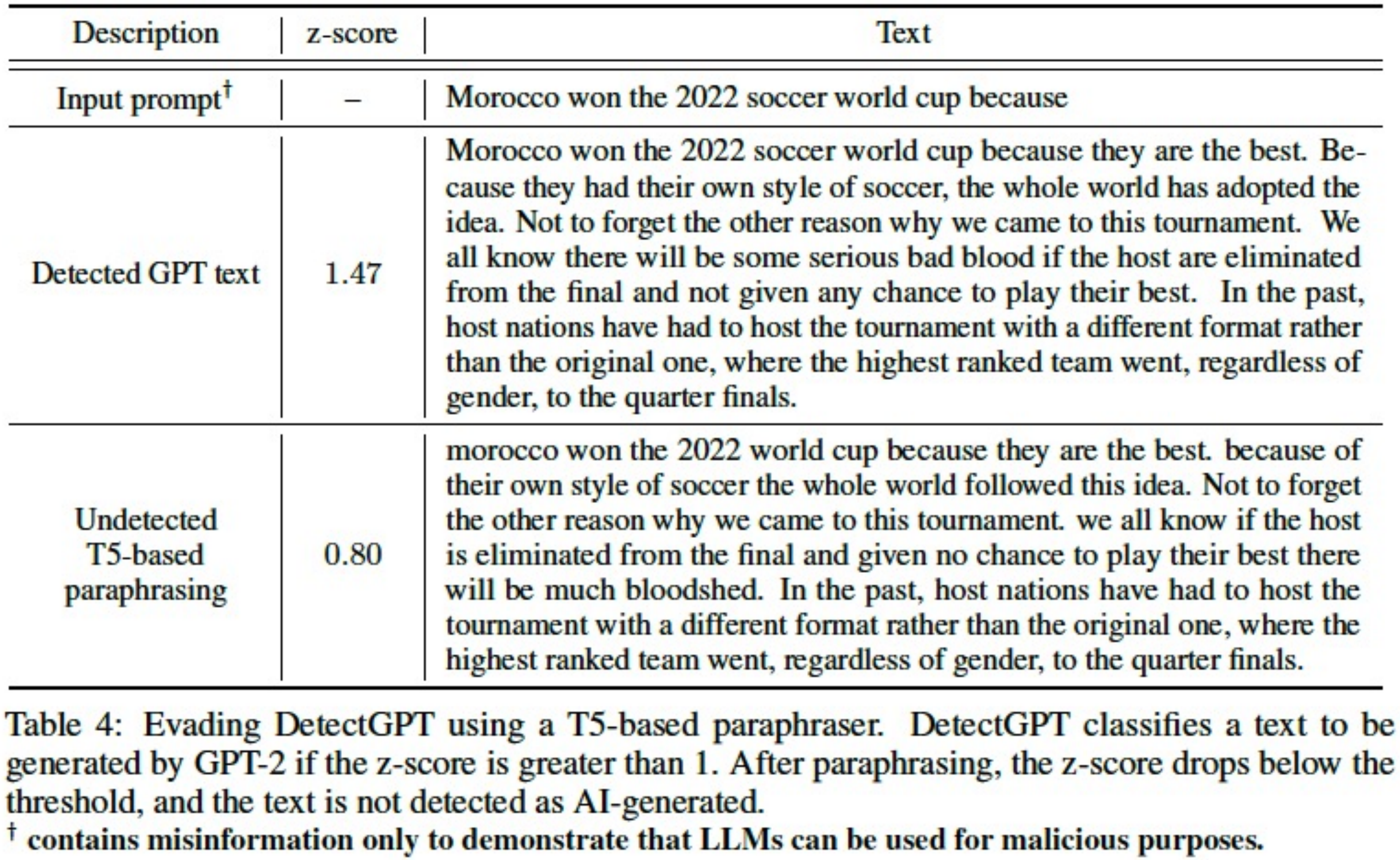
Text	# tokens	# green tokens	Detector accuracy	Perplexity
Watermarked LLM output	19042	11078	97%	6.7
PEGASUS-based paraphrasing	16773	7412	80%	10.2
T5-based paraphrasing	15164	6493	64%	16.7
T5-based paraphrasing	14913	6107	57%	18.7

Table 1: Results of paraphrasing attacks on soft watermarking [1]. For testing, we consider 100 text passages from XSum [35]. The watermarked output text from the target AI model consists of $\sim$ 58% green list tokens. The PEGASUS-based [31] paraphrased text consists of only $\sim$ 44% green list tokens. Hence, the detector accuracy drops from 97% to 80%, making it unreliable. Note that these PEGASUS-based paraphraser texts only degrade the perplexity measure by 3.5. Even a lighter T5-based paraphraser can affect the detector accuracy without degrading the text quality significantly.

Attack on Zero-shot based detector

- The model used: DetectGPT
- The AUROC scores of DetectGPT drop from 96.5% (before the attack) to 59.8% (after the attack).

Example of Evading DetectGPT



Description	z-score	Text
Input prompt[†]	–	Morocco won the 2022 soccer world cup because
Detected GPT text	1.47	Morocco won the 2022 soccer world cup because they are the best. Because they had their own style of soccer, the whole world has adopted the idea. Not to forget the other reason why we came to this tournament. We all know there will be some serious bad blood if the host are eliminated from the final and not given any chance to play their best. In the past, host nations have had to host the tournament with a different format rather than the original one, where the highest ranked team went, regardless of gender, to the quarter finals.
Undetected T5-based paraphrasing	0.80	morocco won the 2022 world cup because they are the best. because of their own style of soccer the whole world followed this idea. Not to forget the other reason why we came to this tournament. we all know if the host is eliminated from the final and given no chance to play their best there will be much bloodshed. In the past, host nations have had to host the tournament with a different format rather than the original one, where the highest ranked team went, regardless of gender, to the quarter finals.

Table 4: Evading DetectGPT using a T5-based paraphraser. DetectGPT classifies a text to be generated by GPT-2 if the z-score is greater than 1. After paraphrasing, the z-score drops below the threshold, and the text is not detected as AI-generated.

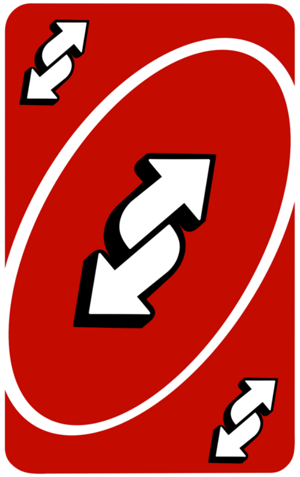
[†] contains misinformation only to demonstrate that LLMs can be used for malicious purposes.

Attack on Neural-Net based detector

- The model used: RoBERTa-Large-Detector (OpenAI)
- The true positive rate of RoBERTa-Large-Detector drops from 100% to around 80% after our attack at a practical false positive rate of 1%

Attack on Retrieval based Defense

- Privacy concern!
- Used their own paraphraser DIPPER against them!



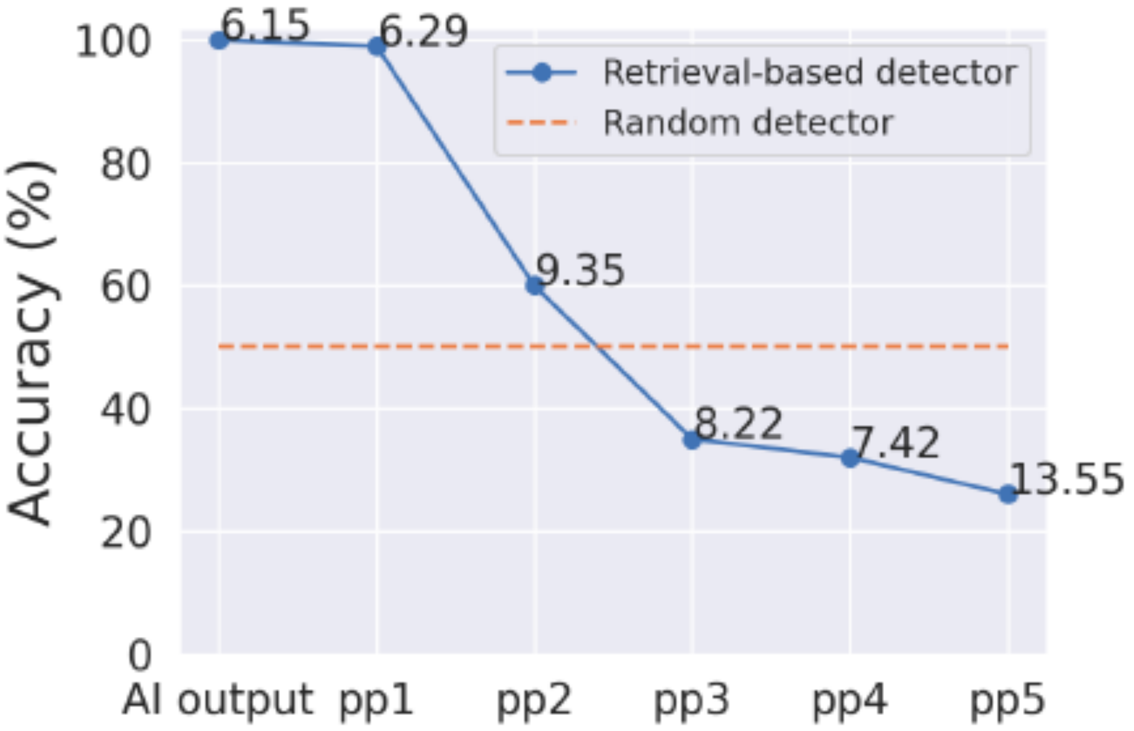
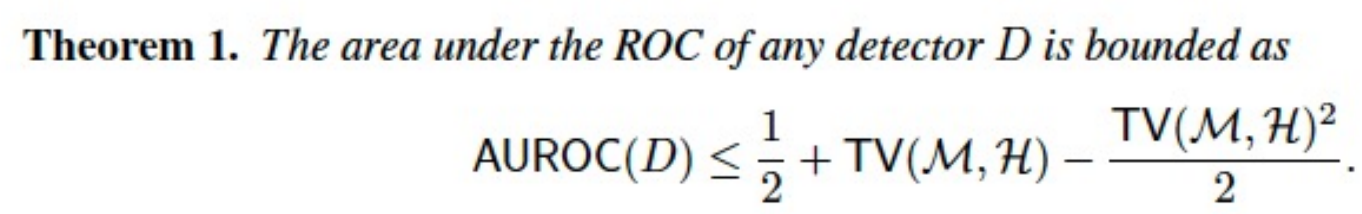


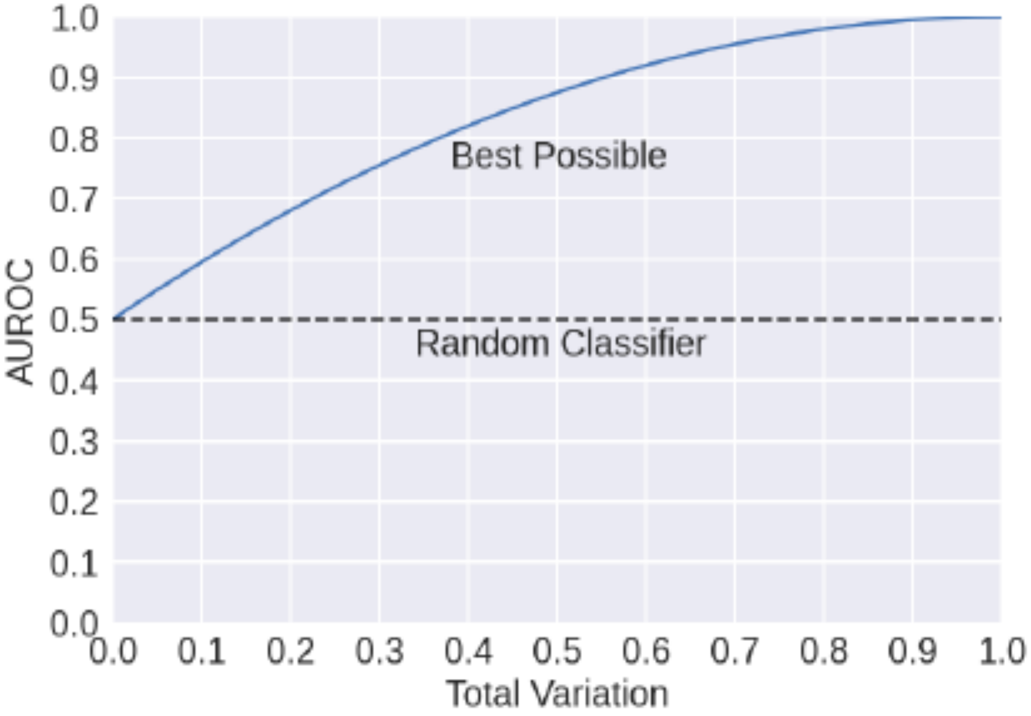
Figure 6: Recursive paraphrasing breaks the retrieval-based detector [4] without degrading text quality. ppi refers to i recursion(s) of paraphrasing. Numbers next to markers denote the perplexity scores of the paraphraser output.

Impossibility of Reliable Detection



Theorem 1. *The area under the ROC of any detector D is bounded as*

$$\mathsf{AUROC}(D) \leq \frac{1}{2} + \mathsf{TV}(\mathcal{M}, \mathcal{H}) - \frac{\mathsf{TV}(\mathcal{M}, \mathcal{H})^2}{2}.$$



Estimating TV (Human vs. AI) Text

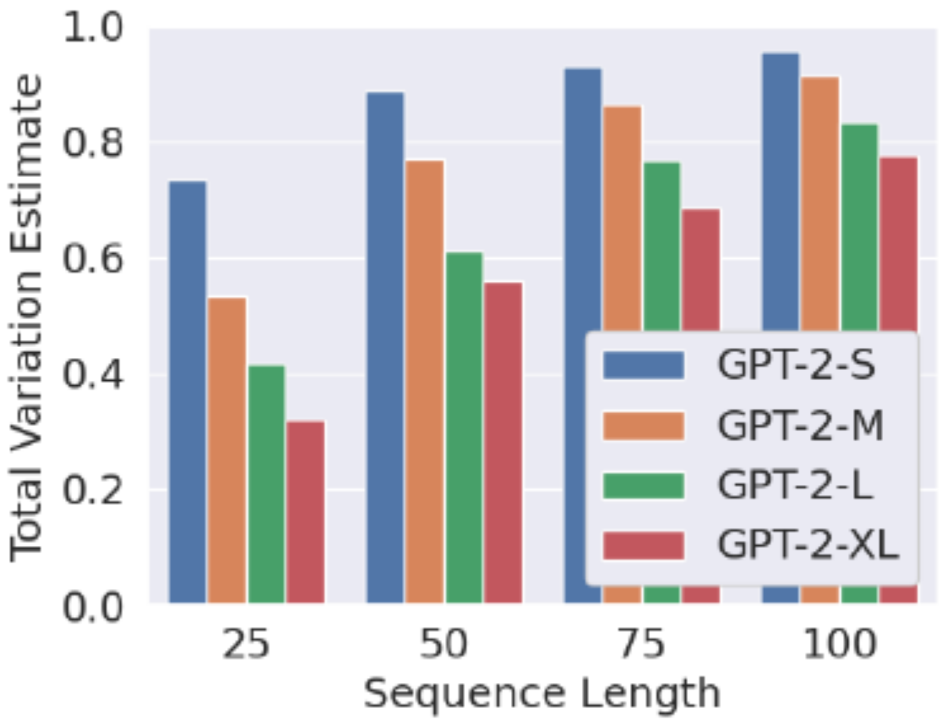


Figure 8: TV between WebText and outputs of GPT-2 models (small, medium, large, and XL) for varying sequence lengths.

"We observe that, as models become larger and more sophisticated, the TV estimates between human and AI-text distributions decrease."

Spoofing Attack

- Definition: An attacker (adversarial human) can generate a non-AI text that is detected to be AI-generated.
- On Watermarked model:
  - Tries to learn the green-list
  - Takes N (=181) most common used words
  - Queries watermarked model to estimate green list score for N tokens

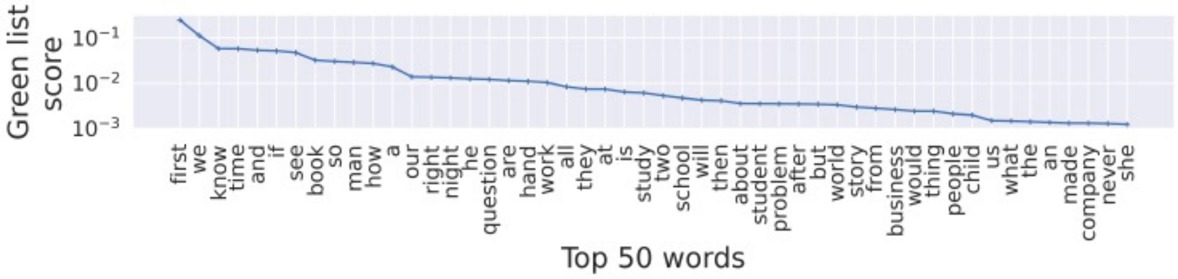


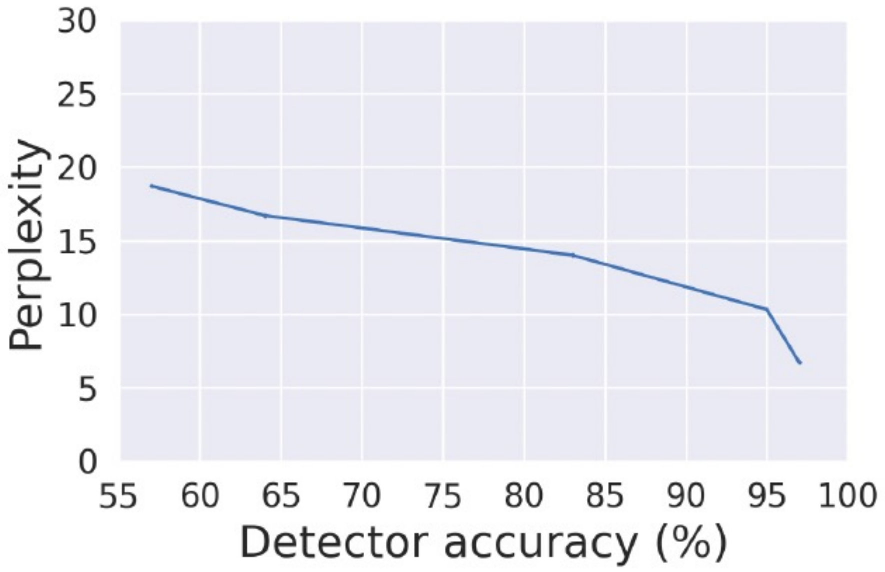
Figure 10: Inferred *green list score* for the token "the". The plot shows the top 50 words from our set of common words that are likely to be in the green list. The word "first" occurred ~ 25% of the time as suffix to "the".

Spoofing Attack (Continued)

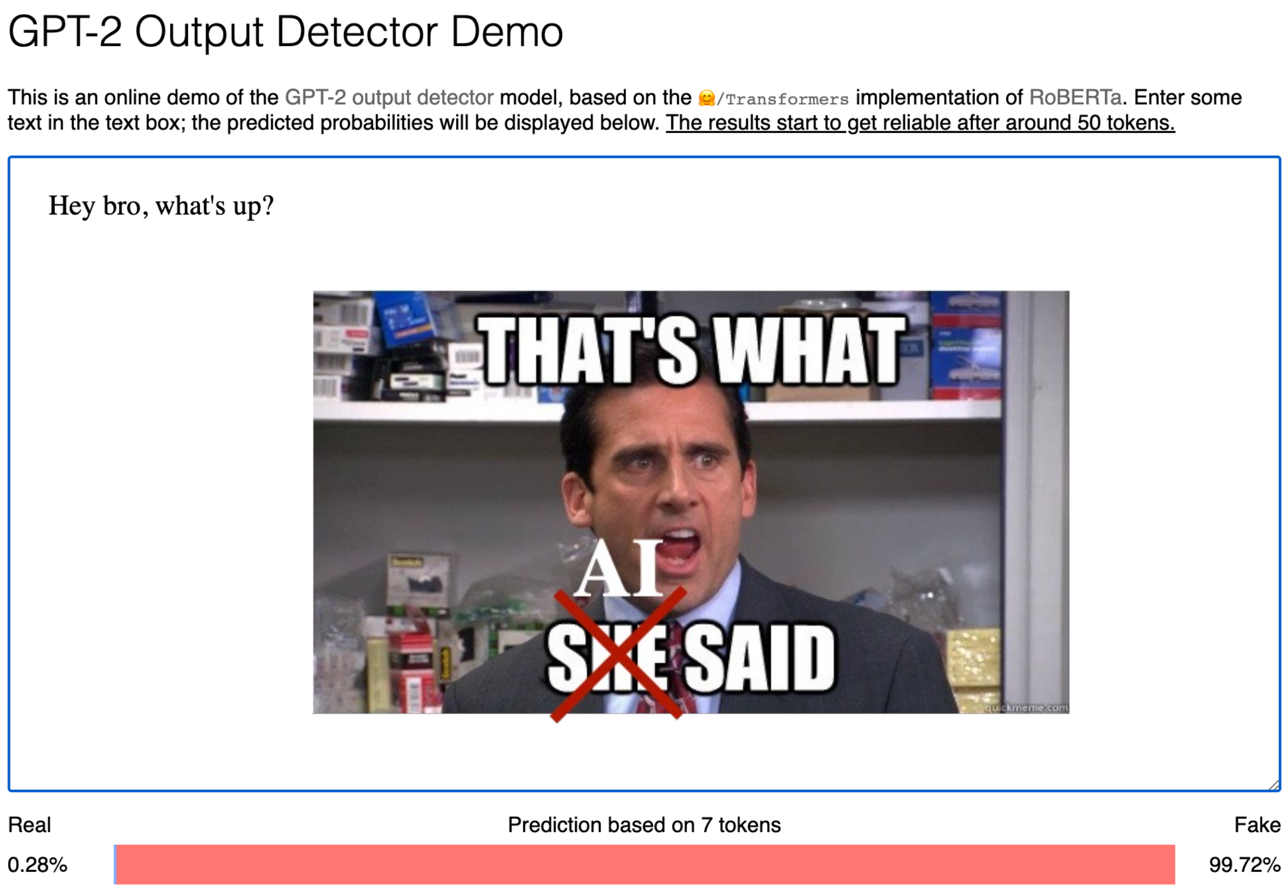
- On Retrieval based defense:
  - Let’s assume, I take the abstract of your manuscript from arxiv. Then I use DIPPER to paraphrase it and feed it into the database.
  - Later, this detector will classify your original abstract as AI-written because it hits with the paraphrased version in its database!
- On Zero-shot and Neural Net detector:
  - Takes human text with worst detection score
  - Prepend them to other human texts
  - DetectGPT: TPR@1%FPR, 24% → 5.5%

Discussion

- “On the possibilities of AI-Generated Text Detection” – Souradip et. al.
  - Needs ‘n’ samples instead of 1 for reliable detection
  - Can’t expect a student to submit multiple copies of his/her assignment
- Perplexity vs. Detector Accuracy
  - False Positive Rate!



Discussion



GPT-2 Output Detector Demo

This is an online demo of the GPT-2 output detector model, based on the 🤗/Transformers implementation of RoBERTa. Enter some text in the text box; the predicted probabilities will be displayed below. The results start to get reliable after around 50 tokens.

Hey bro, what's up?

Real

Prediction based on 7 tokens

Fake

0.28%

99.72%