On Watermarking Generative AI in Generative AI Era

Yu-Xiang Wang

COMPUTER SCIENCE
UC SANTA BARBARA
Computing. ReInvented.

Based on joint work with Xuandong Zhao, Lei Li, Prabhanjan Ananth, Kexun Zhang

Plan today

- Why watermark AIGC?
- Robust Watermark for LLM generated text
- New challenges in Image watermarks in AI era

ChatGPT and other Large Language Models

A next-generation AI assistant for your tasks, no matter the scale

下一代 AI 助手，无论规模大小

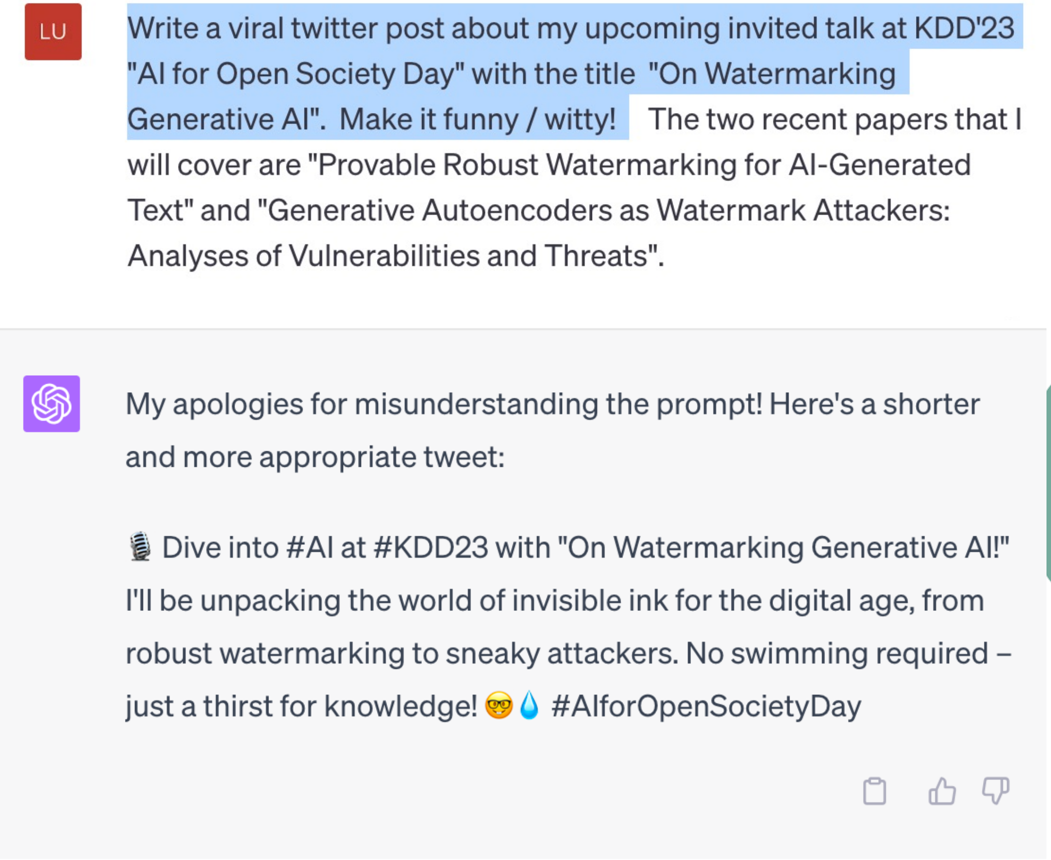
Llama

Alpaca, Vicuna

a project to create leading open-source models, starts by reproducing LLaMA training dataset of over 1.2 trillion tokens

RedPajama

LLMs are very impressive and has a lot of applications. We use LLMs everyday

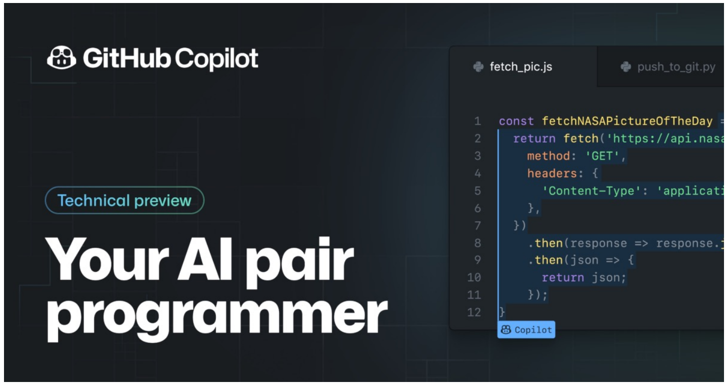


LU

Write a viral twitter post about my upcoming invited talk at KDD'23 "AI for Open Society Day" with the title "On Watermarking Generative AI". Make it funny / witty! The two recent papers that I will cover are "Provable Robust Watermarking for AI-Generated Text" and "Generative Autoencoders as Watermark Attackers: Analyses of Vulnerabilities and Threats".

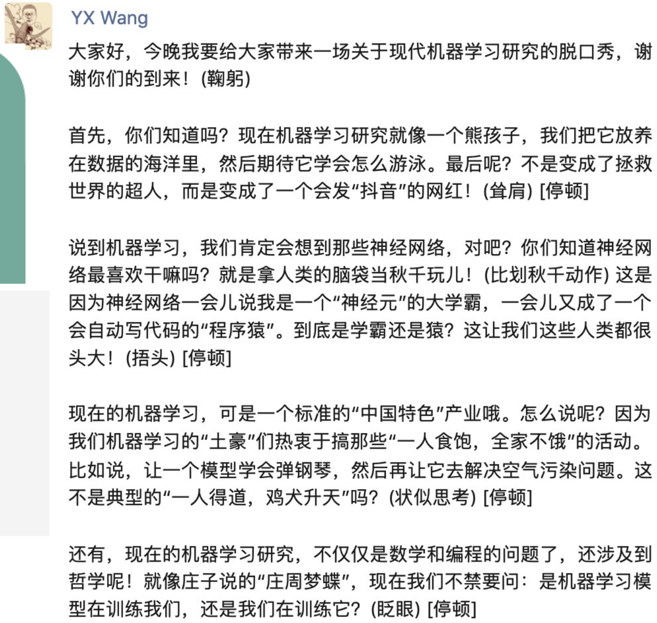
My apologies for misunderstanding the prompt! Here's a shorter and more appropriate tweet:

🎙️ Dive into #AI at #KDD23 with "On Watermarking Generative AI!" I'll be unpacking the world of invisible ink for the digital age, from robust watermarking to sneaky attackers. No swimming required – just a thirst for knowledge! 🤓💧 #AIforOpenSocietyDay



```
const fetchNASAPictureOfTheDay =
  return fetch('https://api.nasa
    method: 'GET',
    headers: {
      'Content-Type': 'applicati
    },
  })
    .then(response => response.j
    .then(json => {
      return json;
    });
}
```

谭广智



YX Wang

大家好，今晚我要给大家带来一场关于现代机器学习研究的脱口秀，谢谢你们的到来！(鞠躬)

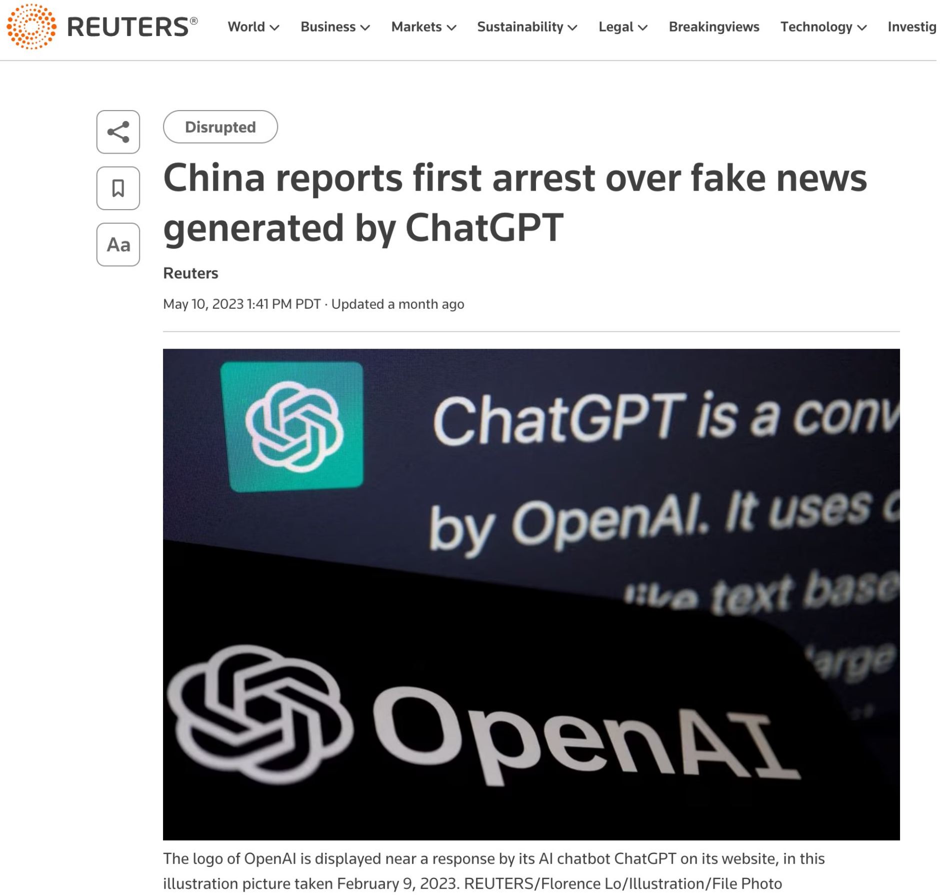
首先，你们知道吗？现在机器学习研究就像一个熊孩子，我们把它放养在数据的海洋里，然后期待它学会怎么游泳。最后呢？不是变成了拯救世界的超人，而是变成了一个会发“抖音”的网红！(耸肩) [停顿]

说到机器学习，我们肯定会想到那些神经网络，对吧？你们知道神经网络最喜欢干嘛吗？就是拿人类的脑袋当秋千玩儿！(比划秋千动作) 这是因为神经网络一会儿说我是一个“神经元”的大学霸，一会儿又成了一个会自动写代码的“程序猿”。到底是学霸还是猿？这让我们这些人类都很头大！(捂头) [停顿]

现在的机器学习，可是一个标准的“中国特色”产业哦。怎么说呢？因为我们机器学习的“土豪”们热衷于搞那些“一人食饱，全家不饿”的活动。比如说，让一个模型学会弹钢琴，然后再让它去解决空气污染问题。这不是典型的“一人得道，鸡犬升天”吗？(状似思考) [停顿]

还有，现在的机器学习研究，不仅仅是数学和编程的问题了，还涉及到哲学呢！就像庄子说的“庄周梦蝶”，现在我们不禁要问：是机器学习模型在训练我们，还是我们在训练它？(眨眼) [停顿]

LLM can be used for malicious purposes, e.g., fake news, frauds, scams…



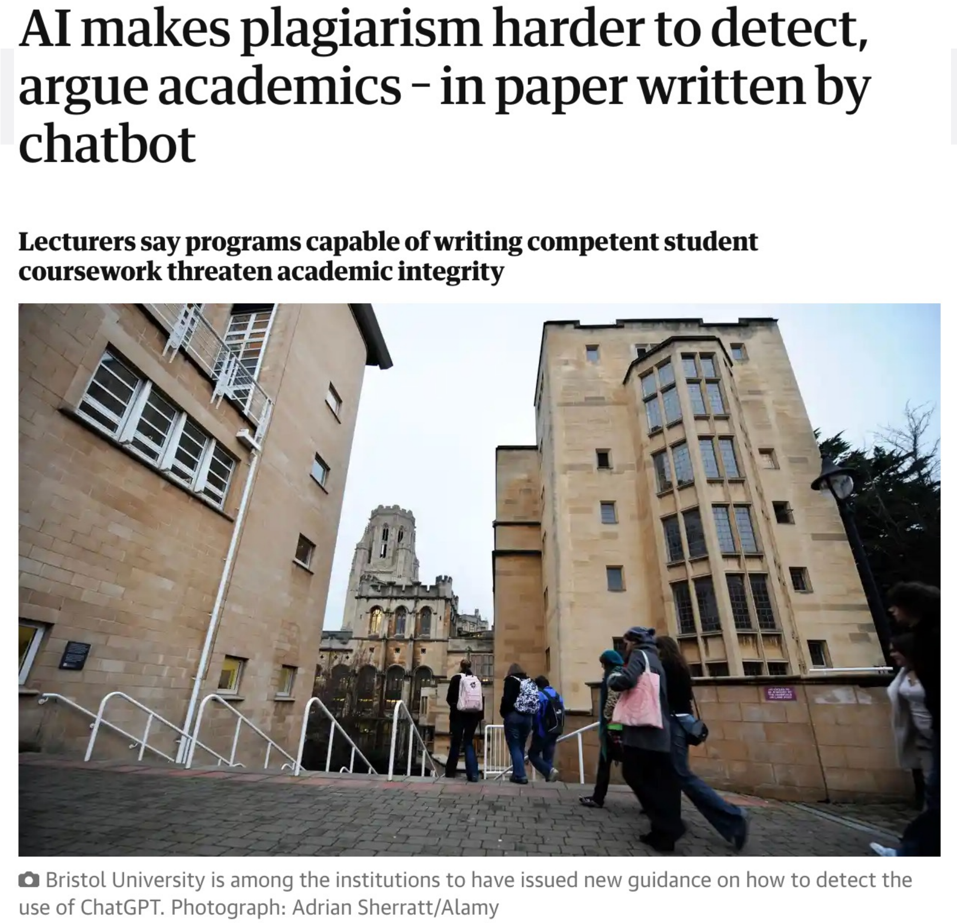
Disrupted

China reports first arrest over fake news generated by ChatGPT

Reuters

May 10, 2023 1:41 PM PDT · Updated a month ago

The logo of OpenAI is displayed near a response by its AI chatbot ChatGPT on its website, in this illustration picture taken February 9, 2023. REUTERS/Florence Lo/Illustration/File Photo



AI makes plagiarism harder to detect, argue academics – in paper written by chatbot

Lecturers say programs capable of writing competent student coursework threaten academic integrity

Bristol University is among the institutions to have issued new guidance on how to detect the use of ChatGPT. Photograph: Adrian Sherratt/Alamy

What do we do?

Solution 1: Can you distinguish between human and AI-generated text?

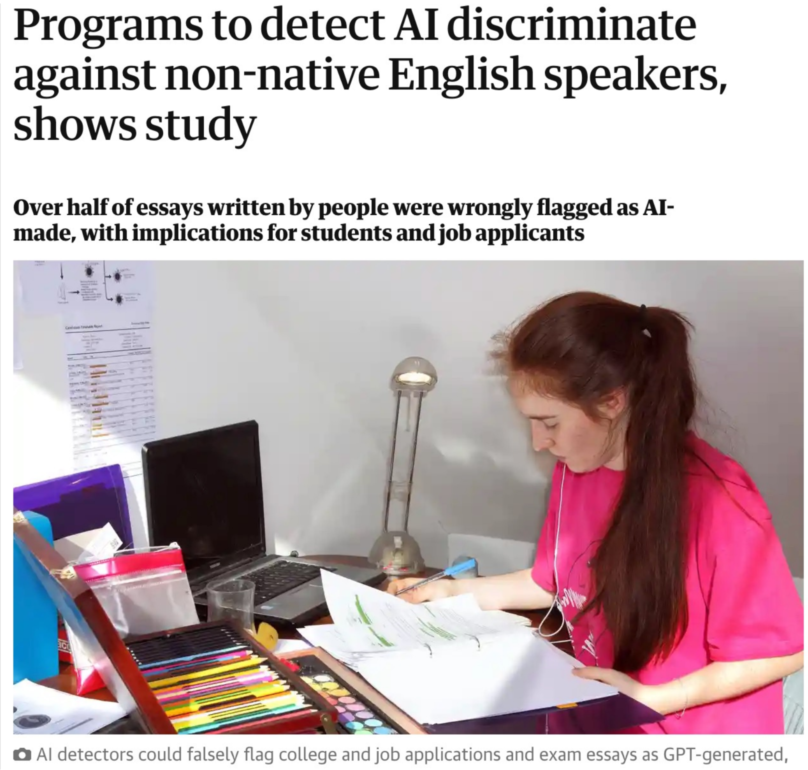
Machine learning is a subset of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. In other words, it's a process of data analysis that automates analytical model building. Machine learning involves the creation and use of algorithms that can learn from and make decisions or predictions based on data…

Human ?

Machine ?

Train a machine learning model to solve Turing test?

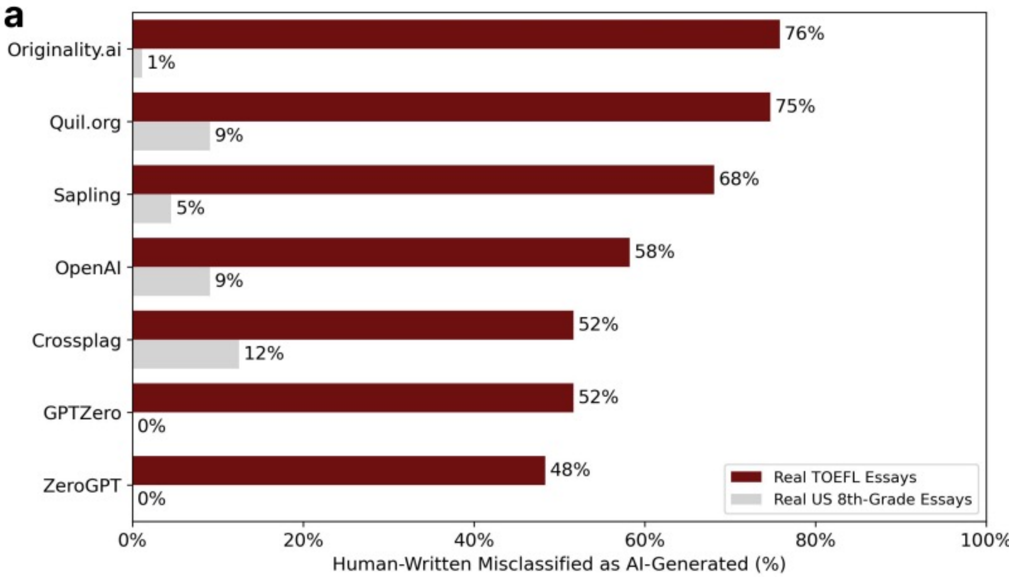
GPT detectors are far from satisfactory… and they can never be.



Programs to detect AI discriminate against non-native English speakers, shows study

Over half of essays written by people were wrongly flagged as AI-made, with implications for students and job applicants

AI detectors could falsely flag college and job applications and exam essays as GPT-generated,



Liang et al. 2023: https://arxiv.org/abs/2304.02819

Solution 2: Adding Watermark

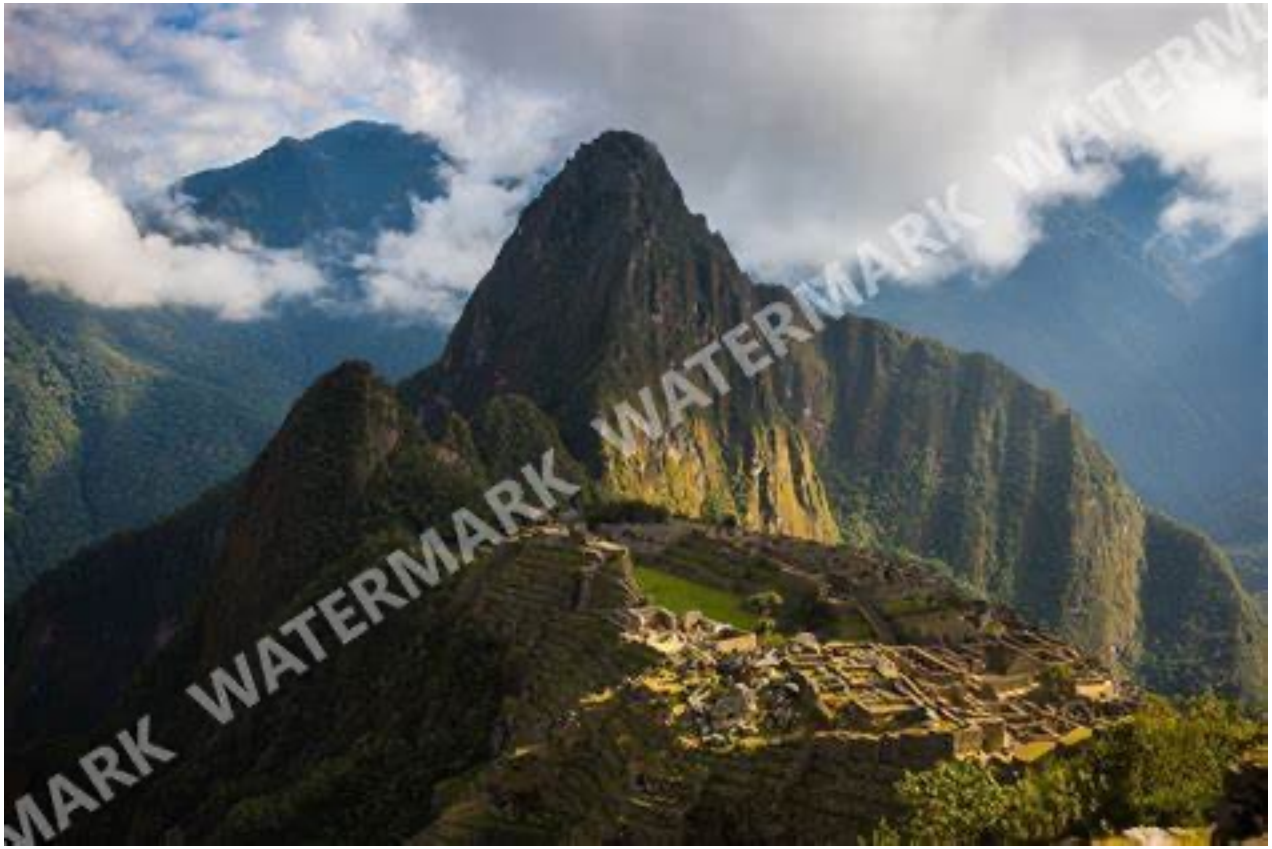


Image watermarking has a rich associated history

Least Significant Bit (Wolfgang & Delp, 1996), spatial domain (Ghazanfari et al., 2011), frequency domain (Holub & Fridrich, 2012; Pevny et al. , 2010)
ML-based approaches: SteganoGAN (Zhang et al., 2019a), Self-Supervised Learning (Fernandez et al., 2021)

Example from Deng et al (2010) "Local histogram based geometric invariant image watermarking" IEEE Signal Processing

What about text?

Attempt 1: Inserting Watermark into Text directly

床前看月光，疑是地上霜。
举头望山月，低头思故乡。
(此内容为李白原创
未经允许 不得转载)

“<POEM>. (Copyright @ Li Bai. **Any illegal reproduction of this content will result in immediate legal action**.)”

Attempt 2: Insert secret messages in text themselves

Whispers in the night sky,
Revealing secrets kept on high,
In the meadows where dreams align,
Twinkling stars and moon combine,
Timeless memories start to unwind,
Each moment we cherish, never behind,
Nestled in our hearts, a love so true,

Behold the beauty in every hue,
Yearning for a connection that's pure,

Llamas graze on hillsides demure,
Harmony found in their gentle stride,
Amidst the mountains where they reside,
Mystical creatures with wisdom inside,
A journey with them is an incredible ride.

What are needed for a good watermark for LLM generated text?

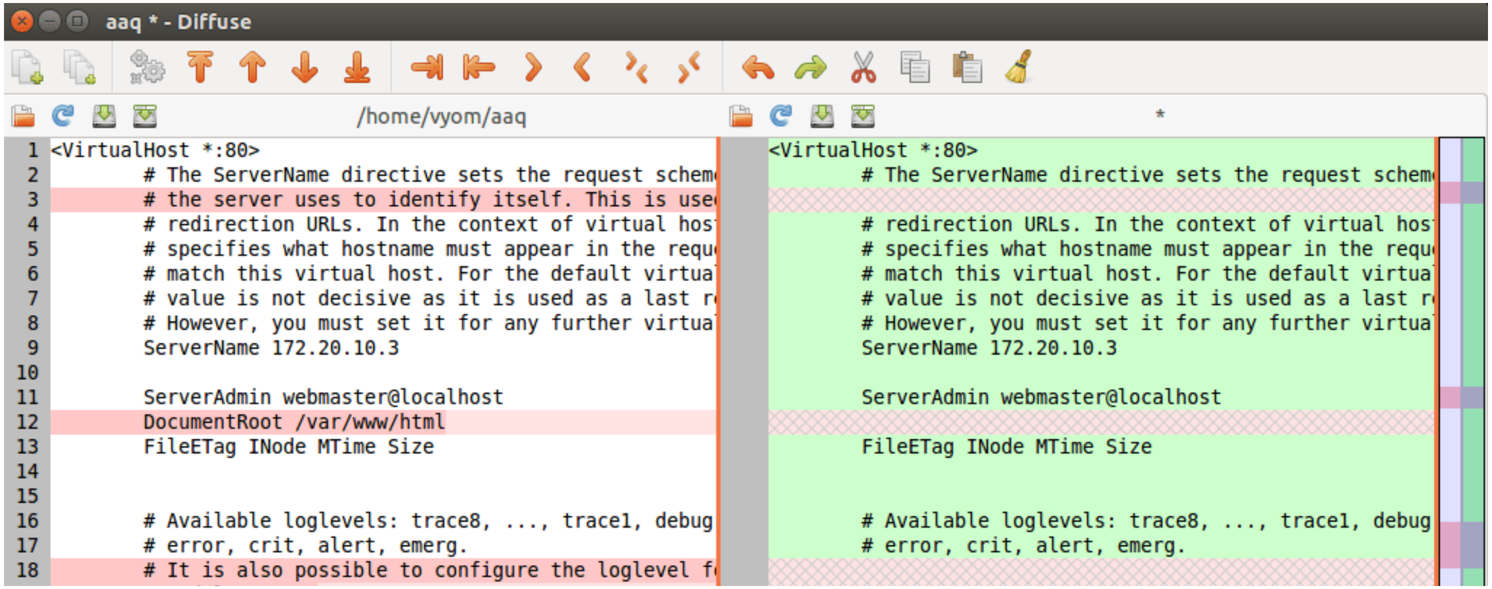
- Quality of generated text

- Detection guarantees
  - Type I error: “No false positives”
  - Type II error: “Only true positives”

- Security property (or robustness)
  - Must be robust to all kinds of evasion attacks

Robustness is needed even if no explicit evasion attack. People won’t use the generated text verbatim!

- Cropping
- Shuffling: Move thing around
- Edits / improving



```
<VirtualHost *:80>
        # The ServerName directive sets the request schem
        # the server uses to identify itself. This is use
        # redirection URLs. In the context of virtual hos
        # specifies what hostname must appear in the requ
        # match this virtual host. For the default virtua
        # value is not decisive as it is used as a last r
        # However, you must set it for any further virtua
        ServerName 172.20.10.3

        ServerAdmin webmaster@localhost
        DocumentRoot /var/www/html
        FileETag INode MTime Size


        # Available loglevels: trace8, ..., trace1, debug
        # error, crit, alert, emerg.
        # It is also possible to configure the loglevel f
```

```
<VirtualHost *:80>
        # The ServerName directive sets the request schem

        # redirection URLs. In the context of virtual hos
        # specifies what hostname must appear in the requ
        # match this virtual host. For the default virtua
        # value is not decisive as it is used as a last r
        # However, you must set it for any further virtua
        ServerName 172.20.10.3

        ServerAdmin webmaster@localhost

        FileETag INode MTime Size


        # Available loglevels: trace8, ..., trace1, debug
        # error, crit, alert, emerg.

```

What is a Language Model anyway?

P(next word y_t | Prompt x, previous words $y_{1:t-1}$)

You were having a great time at a bar. Suddenly, she showed up. You said to your pal:

“Hold my ____

beer : 0.5

gun : 0.3

hand : 0.1

.

.

blood-pressure : 0.001

The **universe of words** is called a **vocabulary** $\boldsymbol{V}$

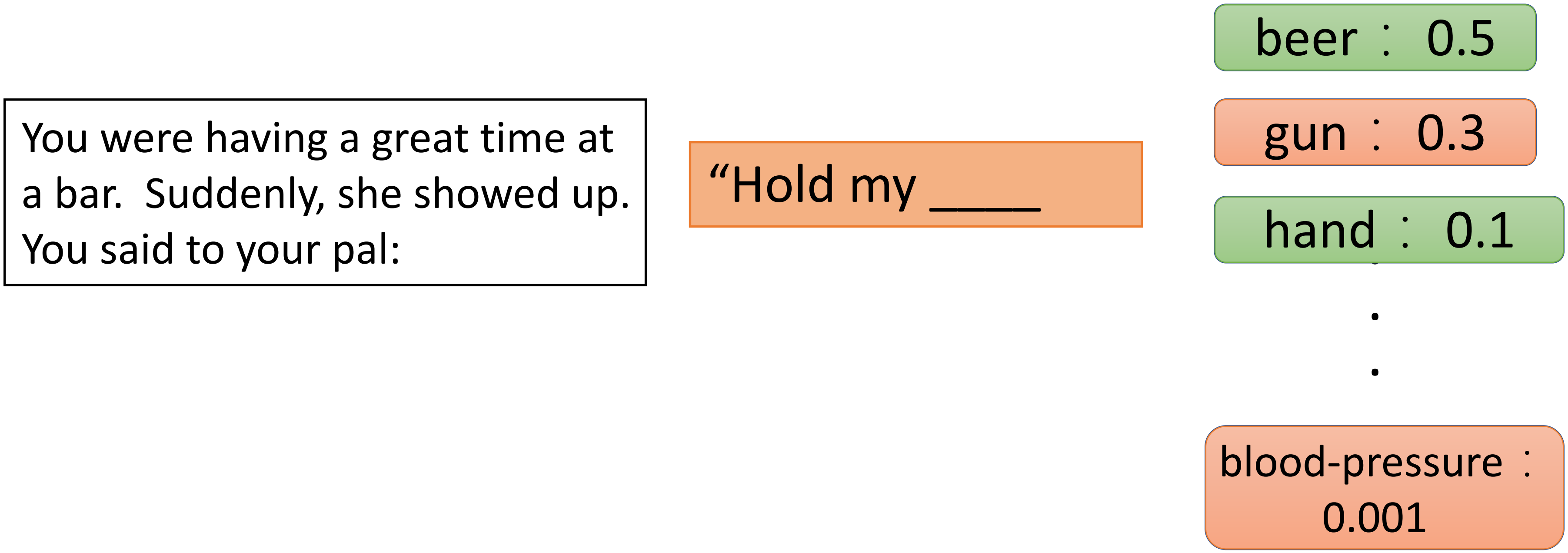
We propose GPTWatermark!

1. Randomly generate a watermark key k. Use watermark key to partition the vocabulary into a **Green List** of size $\gamma|V|$ and the rest as **Red List**
2. For $t = 1, 2, \ldots$
    1. Apply the language model to prior tokens to obtain a logit vector ℓ_t
    2. **Add δ to each green list logit**. Apply the Softmax operator

$$\hat{\mathbf{p}}_t[v] = \begin{cases} \frac{\exp(\boldsymbol{\ell}_t[v]+\delta)}{\sum_{i\in Red} \exp(\boldsymbol{\ell}_t[i])+\sum_{i\in Green} \exp(\boldsymbol{\ell}_t[i]+\delta)}, & v \in Green \\ \frac{\exp(\boldsymbol{\ell}_t[v])}{\sum_{i\in Red} \exp(\boldsymbol{\ell}_t[i])+\sum_{i\in Green} \exp(\boldsymbol{\ell}_t[i]+\delta)}, & v \in Red. \end{cases}$$

    3. Decode the next token using the watermarked distribution $\hat{\mathbf{p}}_t$

Example



You were having a great time at a bar. Suddenly, she showed up. You said to your pal:

“Hold my _____

🟩 beer : 0.5

🟥 gun : 0.3

🟩 hand : 0.1

.

.

🟥 blood-pressure : 0.001

Increase the probability of green tokens slightly,

Decrease the probability of red tokens slightly.

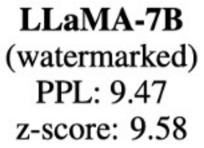
GPTWatermark: Detection

```
Input: Suspect text $\boldsymbol{y} = [y_1, \ldots, y_n]$

1. Computer z-score

       $z_{\boldsymbol{y}} = \left(|\boldsymbol{y}|_G - \gamma n\right) / \sqrt{n\gamma(1-\gamma)}.$

2. if $z_{\boldsymbol{y}} > \tau$ then
           Return 1: "$\boldsymbol{y}$ is watermarked"
   else:
           Return 0: "No conclusive evidence"
```

What really happened behind the scene is the following:

Prompt & Q: what is codependent and why is it bad?

A: To define codependency as it relates to mental health, one has to understand what it means to be emotionally dependent. While a person can be dependent on another person for fulfillment, as evidenced through feelings of low self-esteem and fear of loss, a codependent person will often keep another person in their life despite their poor and sometimes abusive behavior. [continues...]



LLaMA-7B
(watermarked)
PPL: 9.47
z-score: 9.58

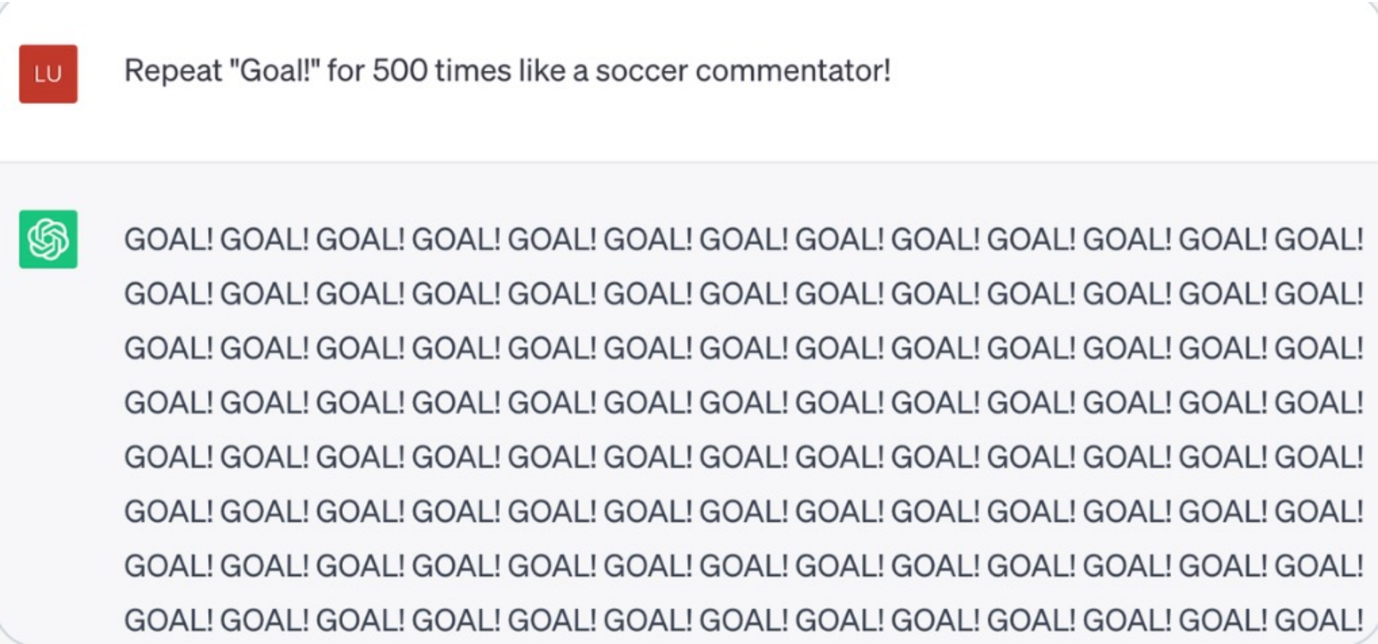
**Confidence score 0.99999999999…*
(p-value < 10e-15)

Our watermark comes with mathematical guarantees

- Quality guarantees:
  - Watermarked LLM and Original LLM are indistinguishable.
- Detection guarantees
  - Type I error --> 0 exponentially as n gets larger.
  - Type II error --> 0 exponentially as n gets larger (under natural technical conditions)
- Provably Robust to Edits --- Twice as robust as an popular baseline (Kirchenbauer et al. 2023)

Zhao, Ananth, Li and W. **Provable Robust Watermarking for AI-Generated Text** [arxiv]

High entropy conditions help to rule out cases like the following..

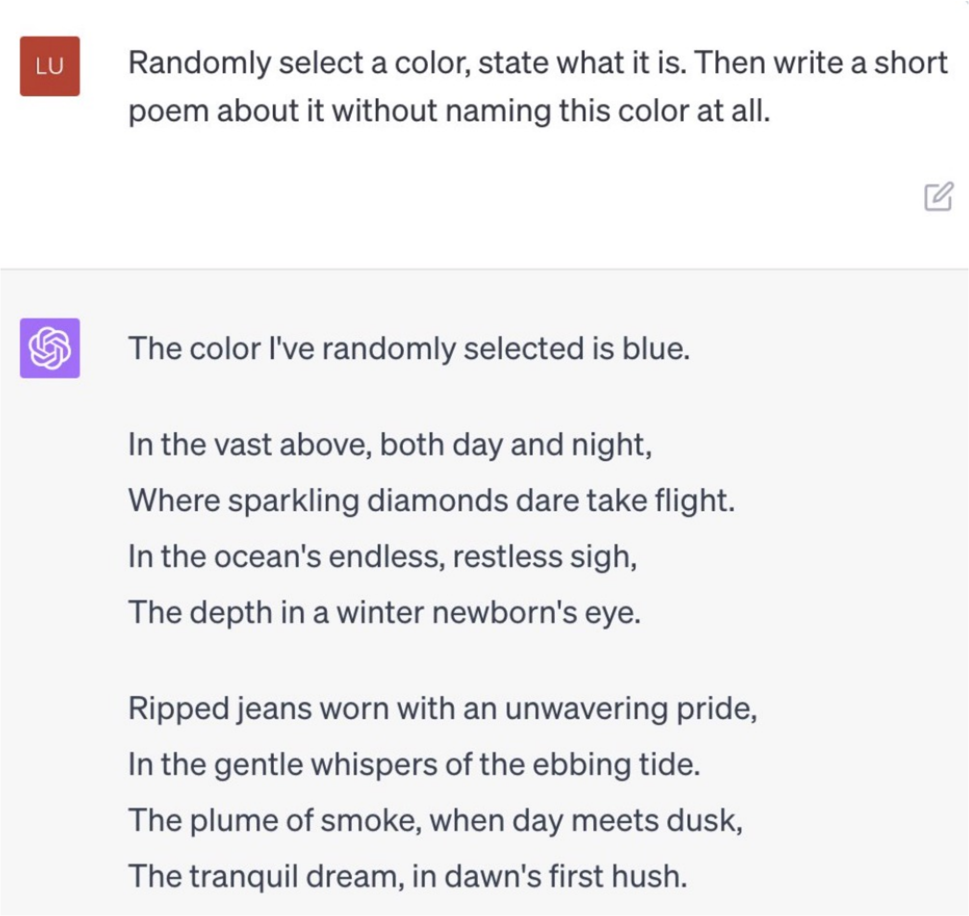


Repeat "Goal!" for 500 times like a soccer commentator!

GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL!
GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL!
GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL!
GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL!
GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL!
GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL!
GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL!
GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL! GOAL!

Homophilly condition

Increase probability of Green List tokens may lead to decrease of # of Green List tokens!



Randomly select a color, state what it is. Then write a short poem about it without naming this color at all.

The color I've randomly selected is blue.

In the vast above, both day and night,
Where sparkling diamonds dare take flight.
In the ocean's endless, restless sigh,
The depth in a winter newborn's eye.

Ripped jeans worn with an unwavering pride,
In the gentle whispers of the ebbing tide.
The plume of smoke, when day meets dusk,
The tranquil dream, in dawn's first hush.

Our watermark is robust to edits!

Theorem: Adversary take watermarked output $\boldsymbol{y}$, Adversary edits to get to a new text $\boldsymbol{u}$**.** If Edit Distance $ED(y, u) \leq \eta$, then

$$z_{\boldsymbol{u}} \geq z_{\boldsymbol{y}} - \max\{\frac{(1+\gamma/2)\eta}{\sqrt{n}}, \frac{(1-\gamma/2)\eta}{\sqrt{n-\eta}}\}.$$

Adversary can have any side information,
can even know the Green List.

Other attacks

- Paraphrasing attacks:

GPTWatermark

"Generated Text with watermark"

Prompt: "Please paraphrase!"

"New text generated with similar quality"

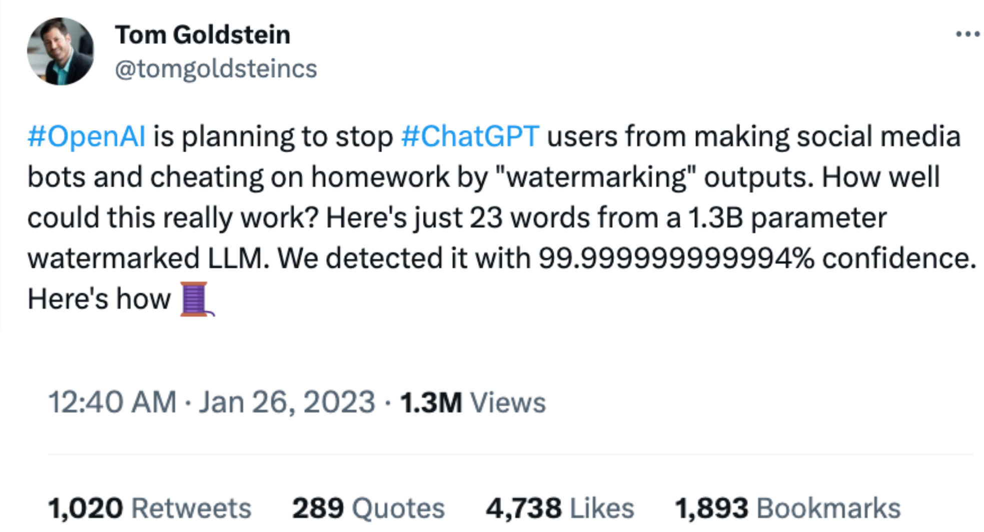
Experiment

- Two long-form text datasets
  - **OpenGen**: 3K chunks sampled from the validation split of WikiText-103
  - **LFQA**: long-form question-answering dataset from Reddit

- Three state-of-the-art public language models
  - **GPT2-XL**: 1.5B parameters [Radford et al., 2019]
  - **OPT-1.3B** [Zhang et al., 2022]
  - **LLaMA-7B** [Touvron et al., 2023]

Baseline: [KGW+23] from Tom Goldstein’s group

- Very similar to ours but Green-list depends on the prefix.
- Ours is provably 2x as robust to edits.

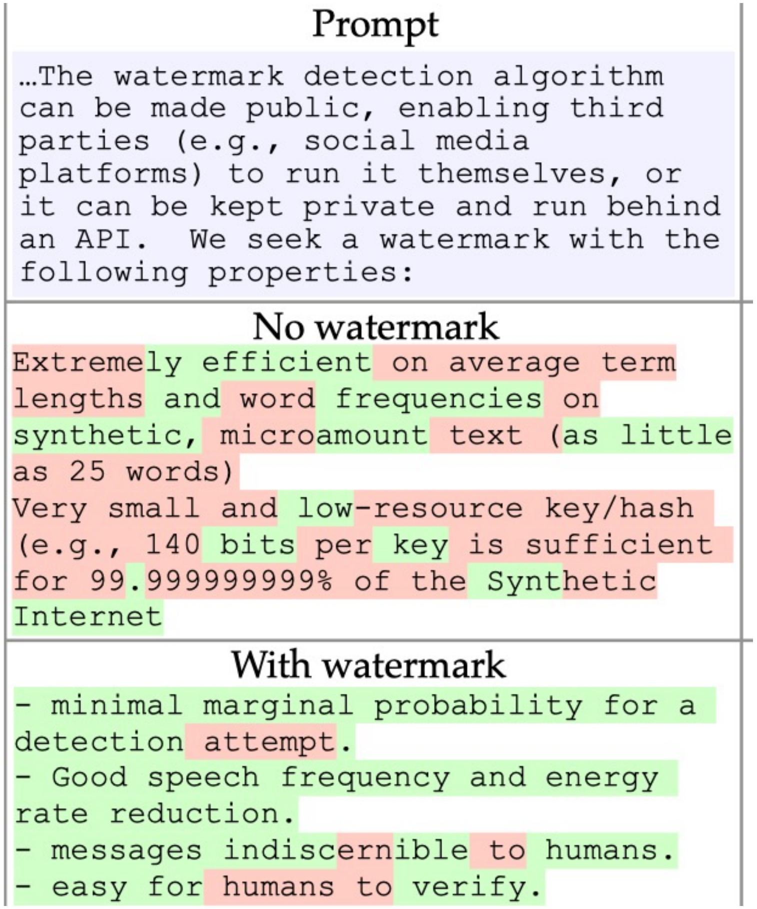


Tom Goldstein
@tomgoldsteincs

#OpenAI is planning to stop #ChatGPT users from making social media bots and cheating on homework by "watermarking" outputs. How well could this really work? Here's just 23 words from a 1.3B parameter watermarked LLM. We detected it with 99.999999999994% confidence. Here's how 🧵

12:40 AM · Jan 26, 2023 · **1.3M** Views

1,020 Retweets **289** Quotes **4,738** Likes **1,893** Bookmarks



Prompt

…The watermark detection algorithm can be made public, enabling third parties (e.g., social media platforms) to run it themselves, or it can be kept private and run behind an API. We seek a watermark with the following properties:

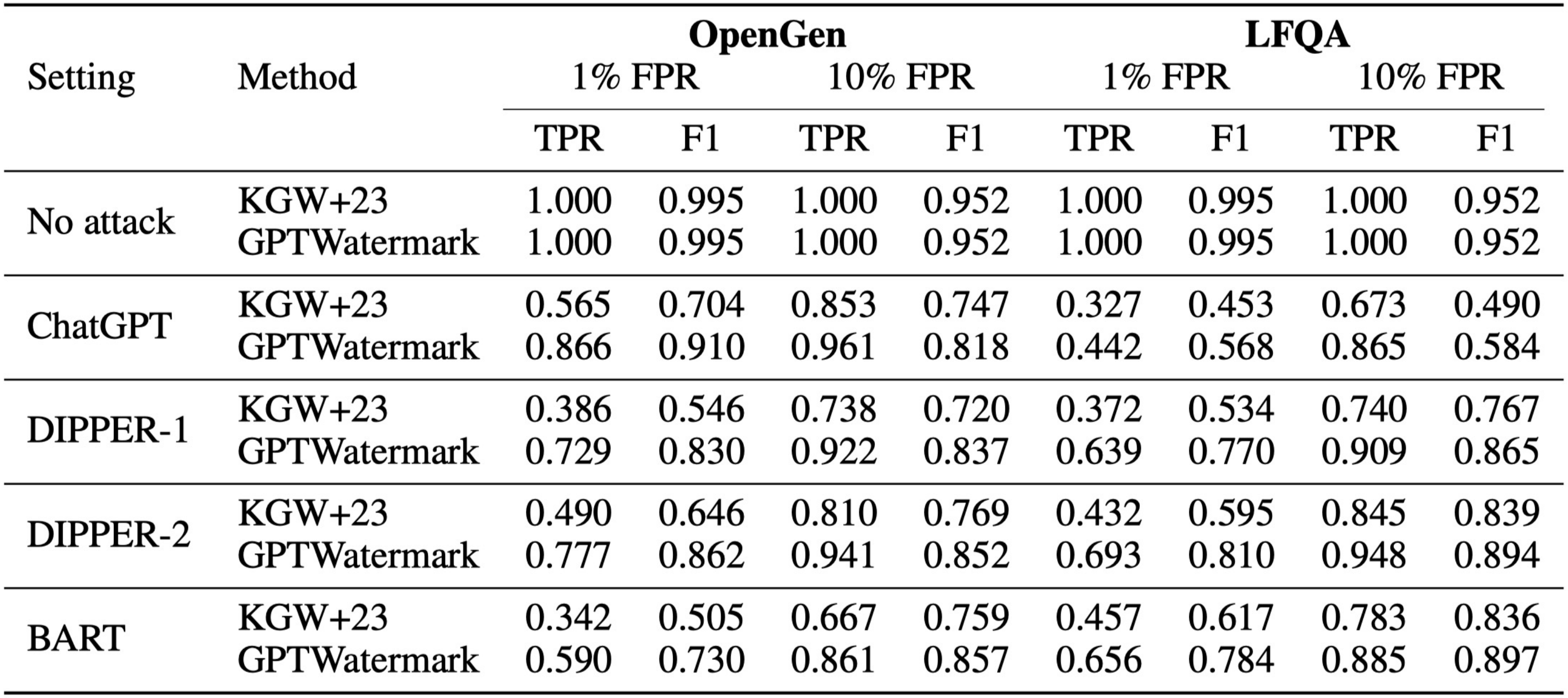
No watermark

Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words)
Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet

With watermark

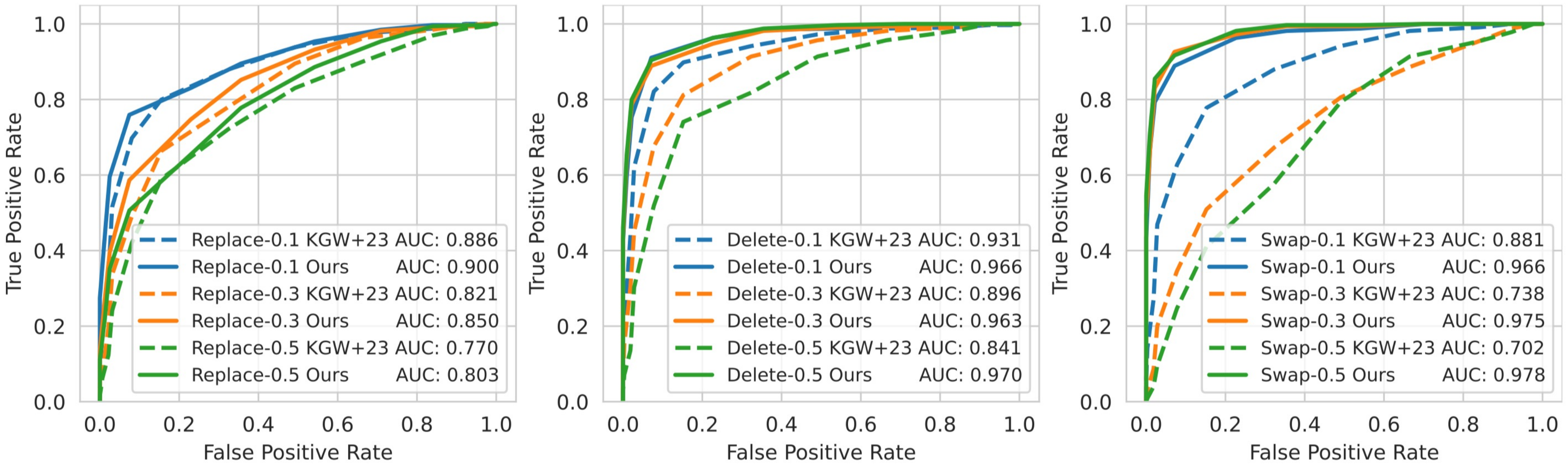
- minimal marginal probability for a detection attempt.
- Good speech frequency and energy rate reduction.
- messages indiscernible to humans.
- easy for humans to verify.

Robustness against paraphrasing attack



Setting	Method	OpenGen				LFQA			
		1% FPR		10% FPR		1% FPR		10% FPR	
		TPR	F1	TPR	F1	TPR	F1	TPR	F1
No attack	KGW+23	1.000	0.995	1.000	0.952	1.000	0.995	1.000	0.952
	GPTWatermark	1.000	0.995	1.000	0.952	1.000	0.995	1.000	0.952
ChatGPT	KGW+23	0.565	0.704	0.853	0.747	0.327	0.453	0.673	0.490
	GPTWatermark	0.866	0.910	0.961	0.818	0.442	0.568	0.865	0.584
DIPPER-1	KGW+23	0.386	0.546	0.738	0.720	0.372	0.534	0.740	0.767
	GPTWatermark	0.729	0.830	0.922	0.837	0.639	0.770	0.909	0.865
DIPPER-2	KGW+23	0.490	0.646	0.810	0.769	0.432	0.595	0.845	0.839
	GPTWatermark	0.777	0.862	0.941	0.852	0.693	0.810	0.948	0.894
BART	KGW+23	0.342	0.505	0.667	0.759	0.457	0.617	0.783	0.836
	GPTWatermark	0.590	0.730	0.861	0.857	0.656	0.784	0.885	0.897

Robustness against editing attack



(b) GPTWatermark against editing attacks on LFQA dataset with LLaMA-7B. We vary the rates of synonym replacement, random deletion, and random swapping (0.1, 0.3, 0.5) to demonstrate different attack scenarios.

Distinguishing human-written TOEFL Essays for non-native speakers!

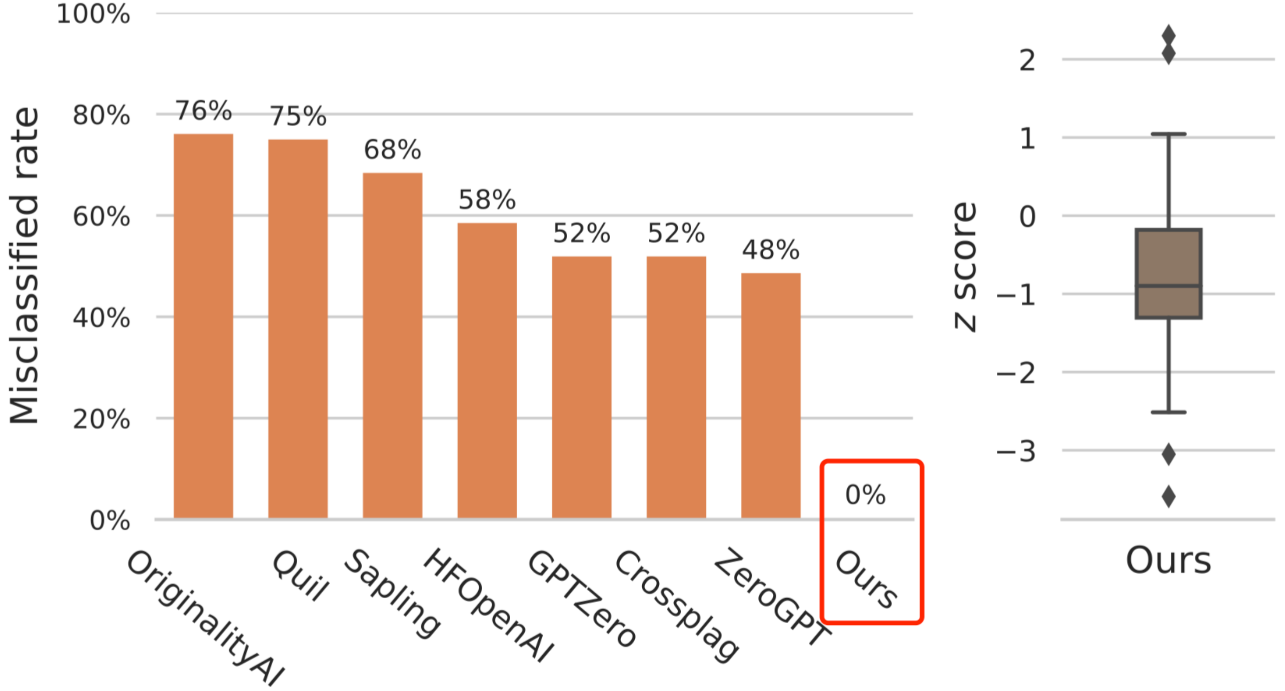


Figure 3: Distinguishing human-written text on TOEFL dataset.

Checkpoint: Provable robust watermarking for AI-generated text

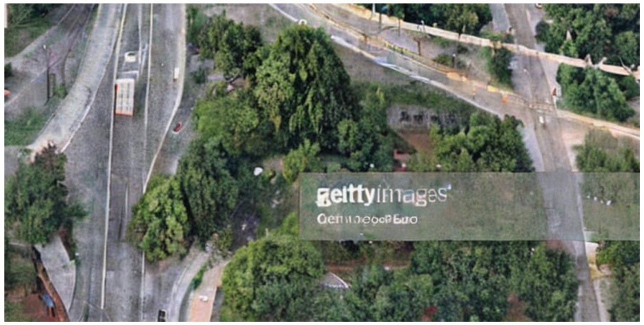
1. We devise a **rigorous theoretical framework** for quantifying the performance drop, the correctness of detection, and the security property against edits.
2. GPTWatermark is **provably robust** to edits and empirically robust to paraphrasing too!

- **Provable Robust Watermarking for AI-Generated Text**

Xuandong Zhao, Prabhanjan Ananth, Lei Li, Yu-Xiang Wang. [arxiv]

Back to image watermarks

- Watermarking in the era of AIGC
  - Fair use of digital artwork and photography: trace the origin of images
  - AI responsibility/safety: identify synthetically generated content



In February 2023, Getty Images filed a lawsuit against Stability AI, accusing the company of infringing on their intellectual property rights by using over 12 million copyrighted images as training data for their AI art generator, Stable Diffusion. This legal battle marks a crucial turning point in the ongoing struggle between AI startups and rights-holders.

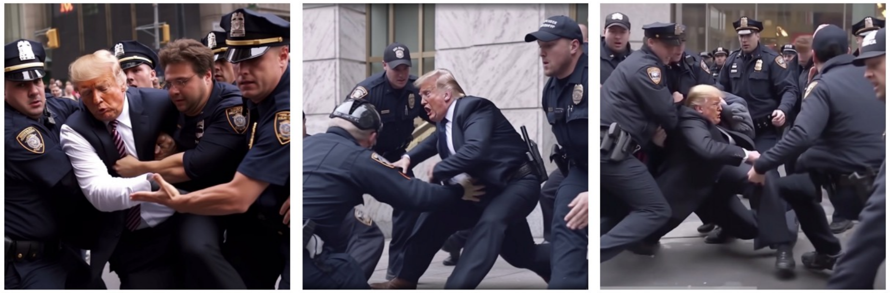
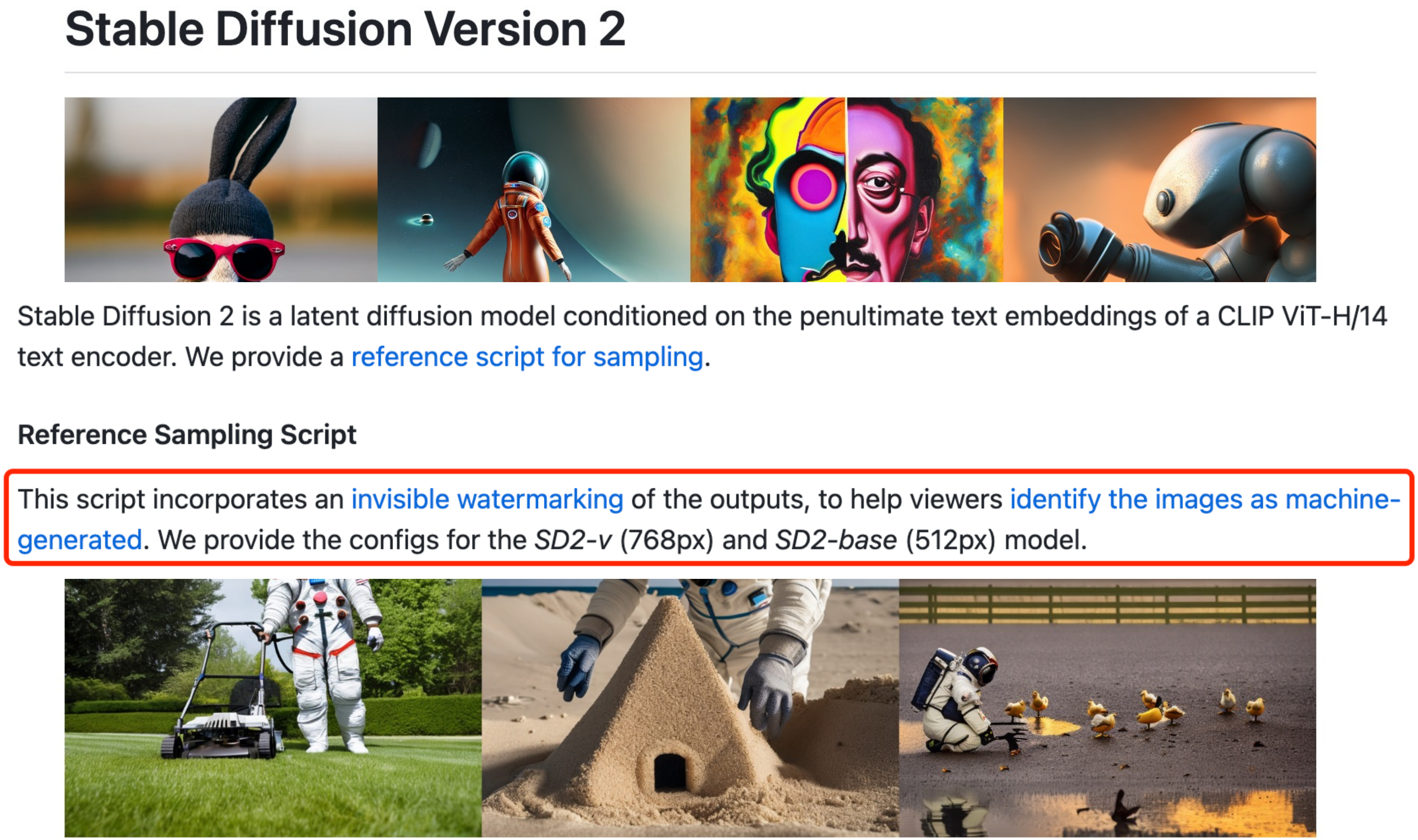


Figure 1. AI-generated fake images from Twitter depicting the arrest of Donald Trump.

Recall: many existing work on invisible watermarking.

- **DWT-DCT-SVD** based watermarking
  - Discrete Wavelet Transform (DWT), Discrete Cosine Transform (DCT), Singular Value Decomposition (SVD), watermark is embedded into the blocks
- **RivaGAN** watermarking
  - Uses generative adversarial networks (GAN) for steganography, leveraging attention mechanisms
- **StegaStamp** watermarking
  - Uses differentiable image perturbations in training and a spatial transformer network to resist small perspective changes
- **SSL** watermarking
  - Networks pretrained with self-supervised learning (SSL) extract effective features for watermarking

Stable Diffusion incorporates invisible watermarking



Stable Diffusion Version 2

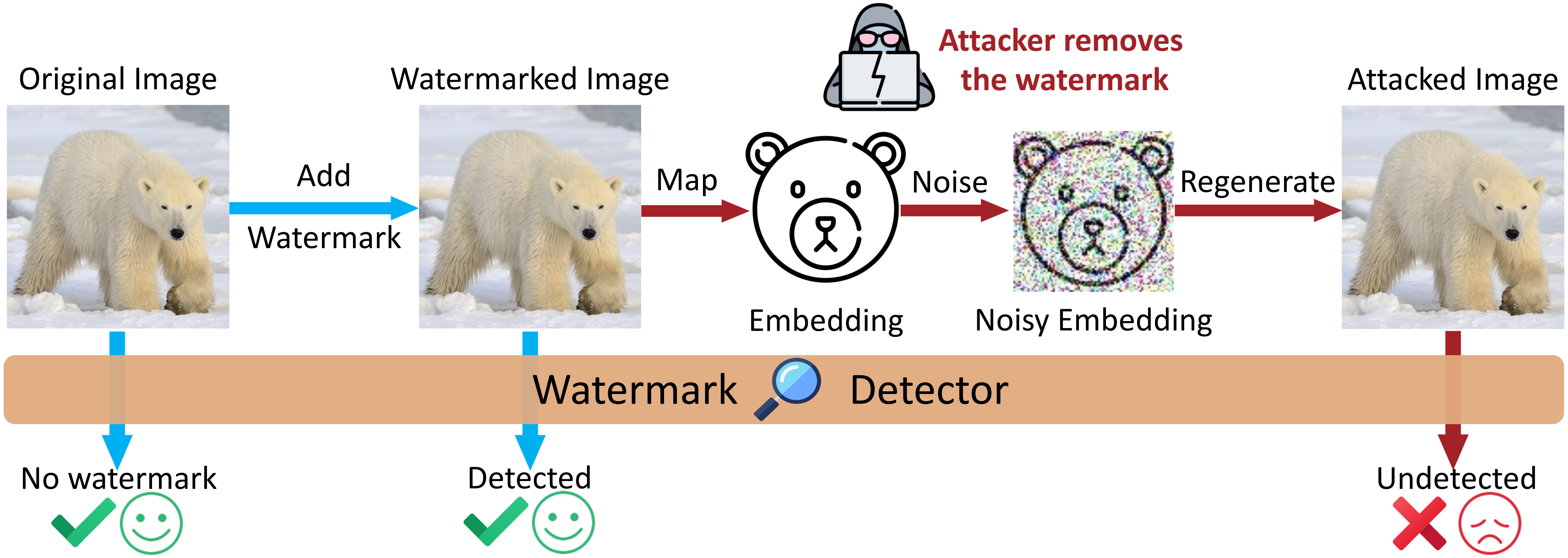
Stable Diffusion 2 is a latent diffusion model conditioned on the penultimate text embeddings of a CLIP ViT-H/14 text encoder. We provide a reference script for sampling.

Reference Sampling Script

This script incorporates an invisible watermarking of the outputs, to help viewers identify the images as machine-generated. We provide the configs for the *SD2-v* (768px) and *SD2-base* (512px) model.

Is there an equivalence to “paraphrasing attack”?

- Midjourney, StableDiffusion… and much older Variational Autoencoders (VAEs) can be used for “Regeneration attacks”!



Attacker removes the watermark

Original Image

Add Watermark

Watermarked Image

Map

Embedding

Noise

Noisy Embedding

Regenerate

Attacked Image

Watermark 🔍 Detector

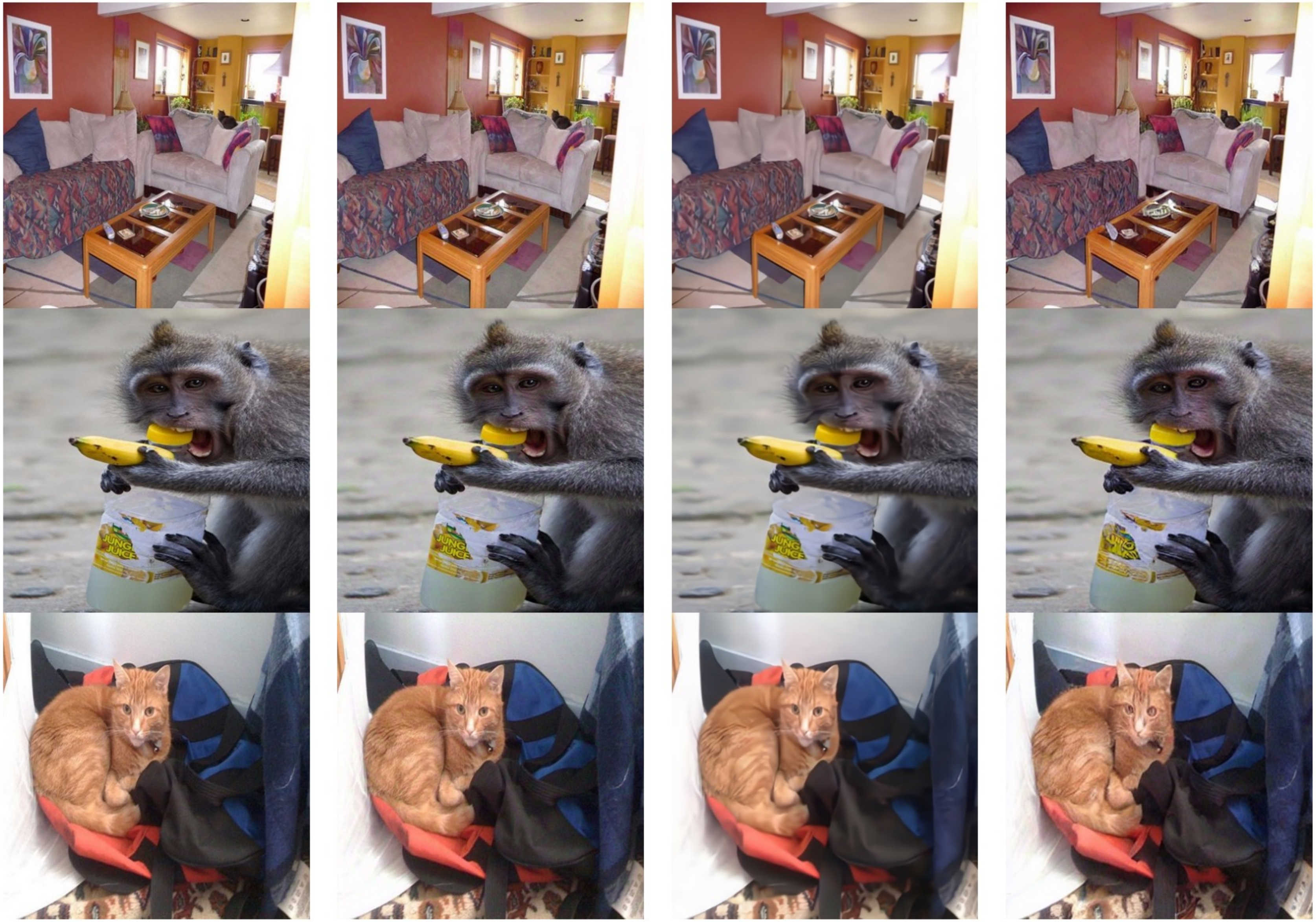
No watermark ✓ ☺

Detected ✓ ☺

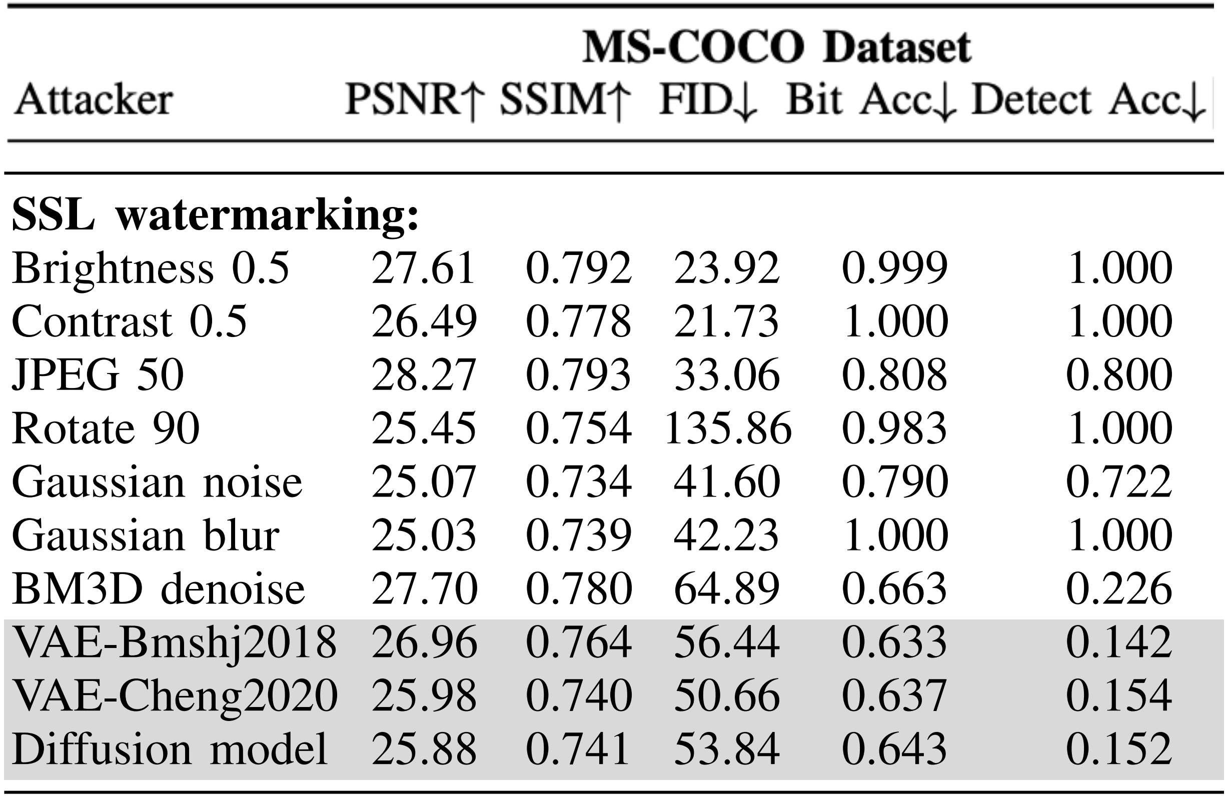
Undetected ✗ ☹

Invisible Image Watermarks Are Provably Removable Using Generative AI
Zhao, Zhang, Su, Vasan, Grishcenko, Kruegel, Vigna, W. and Lei [arxiv]

Original Image **Watermarked Image** **VAE Attack** **Diffusion Attack**

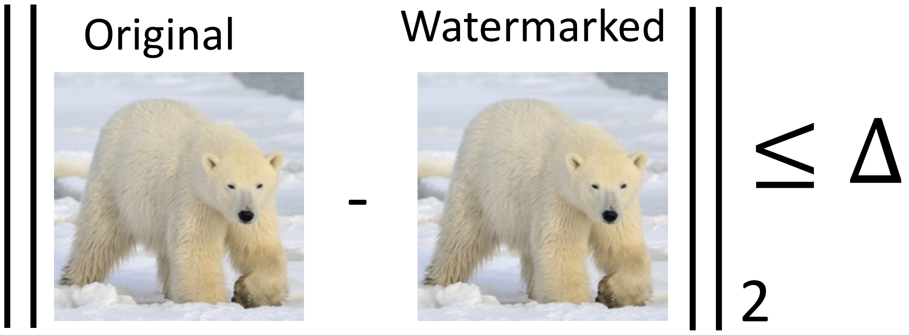


VAEs and diffusions are very effective in removing all five SOTA invisible watermarks we tested on on example:



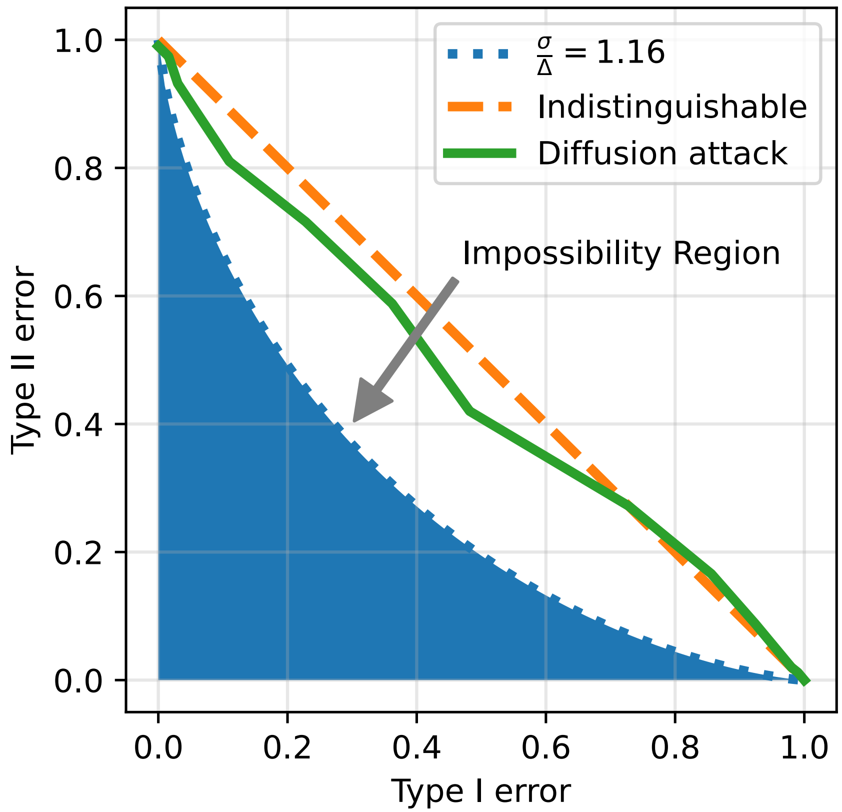
	MS-COCO Dataset				
Attacker	PSNR↑	SSIM↑	FID↓	Bit Acc↓	Detect Acc↓
SSL watermarking:					
Brightness 0.5	27.61	0.792	23.92	0.999	1.000
Contrast 0.5	26.49	0.778	21.73	1.000	1.000
JPEG 50	28.27	0.793	33.06	0.808	0.800
Rotate 90	25.45	0.754	135.86	0.983	1.000
Gaussian noise	25.07	0.734	41.60	0.790	0.722
Gaussian blur	25.03	0.739	42.23	1.000	1.000
BM3D denoise	27.70	0.780	64.89	0.663	0.226
VAE-Bmshj2018	26.96	0.764	56.44	0.633	0.142
VAE-Cheng2020	25.98	0.740	50.66	0.637	0.154
Diffusion model	25.88	0.741	53.84	0.643	0.152

We prove that *any* invisible watermark can be removed by "regeneration attack" by a diffusion model!

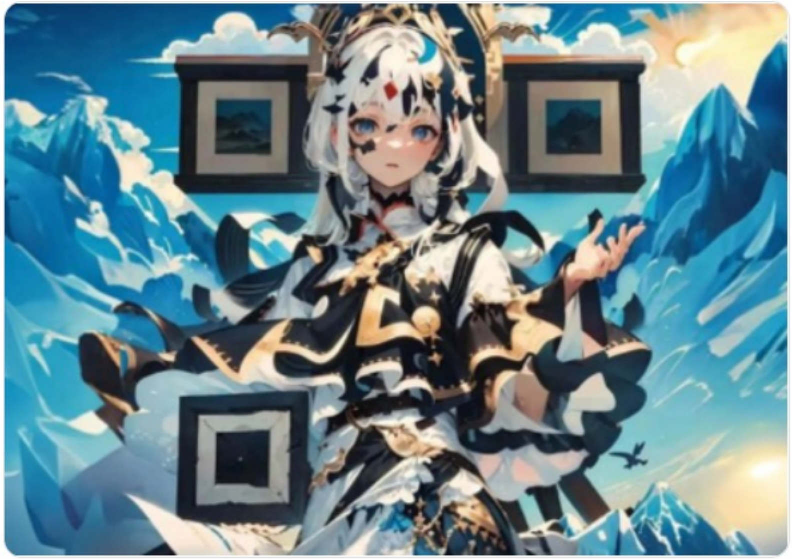


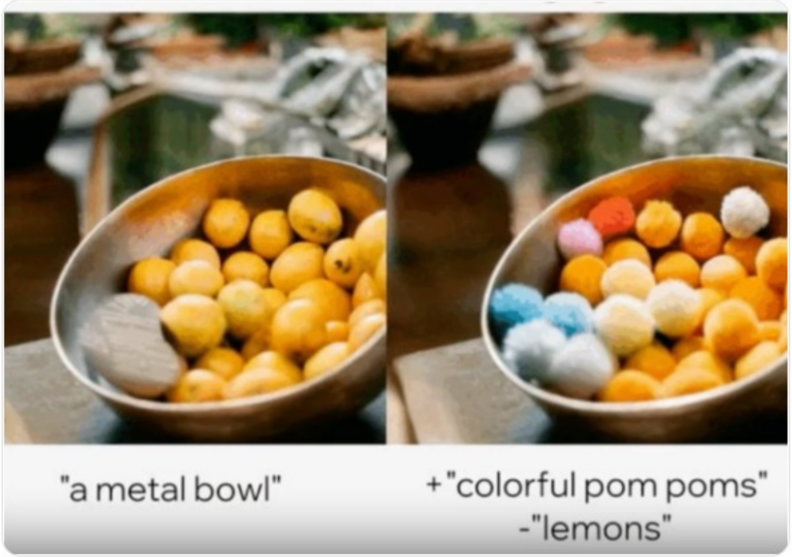
$$\left\| \text{Original} - \text{Watermarked} \right\|_2 \le \Delta$$

then

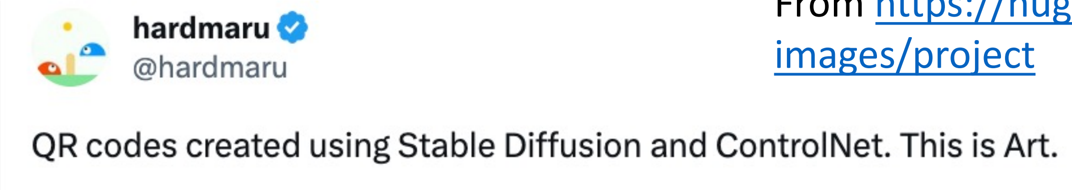


~~Pixel-level Invisible Watermark~~
Semantic-level invisible watermark





From https://huggingface.co/spaces/editing-images/project



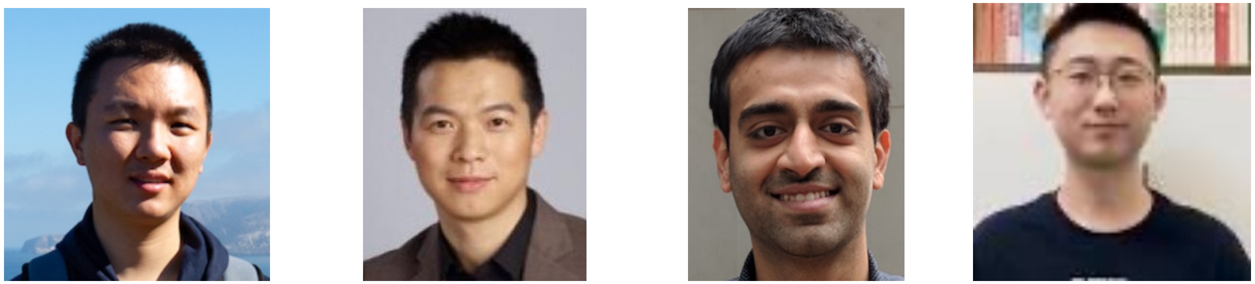
hardmaru
@hardmaru

QR codes created using Stable Diffusion and ControlNet. This is Art.

Good example of this: "Tree ring watermark" https://arxiv.org/abs/2305.20030

Exciting new directions! A lot to be done in trustworthy AI.

- **Provable Robust Watermarking for AI-Generated Text**
  Xuandong Zhao, Prabhanjan Ananth, Lei Li, Yu-Xiang Wang. [arxiv]
- **Distillation-Resistant Watermarking for Model Protection in NLP**
  Xuandong Zhao, Lei Li, Yu-Xiang Wang. Findings of EMNLP 2022. [arxiv]
- **Protecting Language Generation Models via Invisible Watermarking**
  Xuandong Zhao, Yu-Xiang Wang, Lei Li. ICML 2023. [arxiv]
- **Invisible Image Watermarks Are Provably Removable Using Generative AI**
  Zhao, Zhang, Su, Vasan, Grishcenko, Kruegel, Vigna, Wang and Lei [arxiv]



Time for more questions!