

Is It Generated by AI?

Attacks and Robust
Watermarking for Generative AI

Lei Li

January 22, 2024

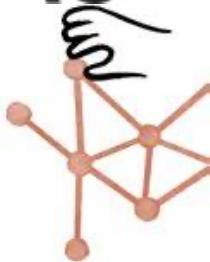
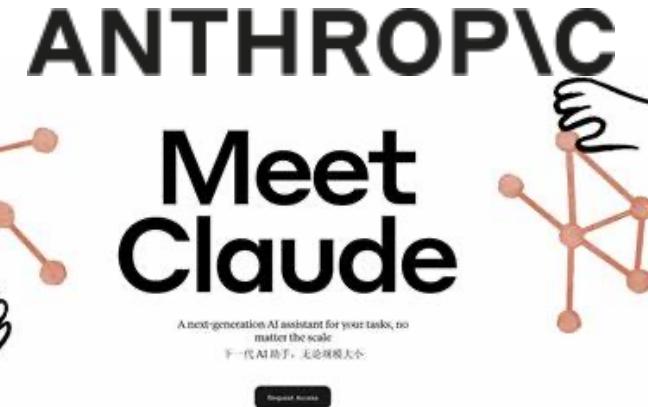
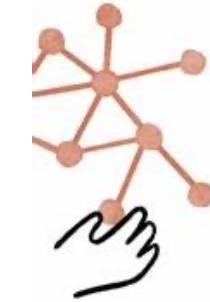
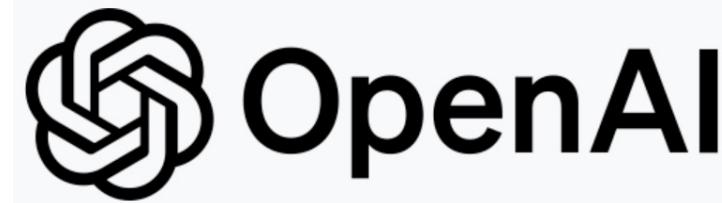


Carnegie Mellon University
Security and Privacy Institute



Carnegie Mellon University
Language Technologies Institute

Large Language Model Products



LLM can

- Translate
- Summarize
- Proof-read and correct
- Explain words
- Create article/email
- Make restaurant suggestions
- Chat with users
- Do Math Calculations
- Answer questions on many subjects
- Suggest names
- Write code

create a viral tweet about my upcoming talk at Carnegie Mellon University (CMU, in Pittsburgh) with the title "Watermark for Large Language Models". Topics include watermark, detecting AI generated text, protecting copyright of models. Make it funny and witty.

 Hey yinzers at @CarnegieMellon!  Get ready to turn "yinzpiration" into action with my talk on "Empowering Responsible Use of Large Language Models"!  

We're gonna tackle AI like it's the Steelers vs. the Ravens , with tricks like watermarking (no, not on Primanti's sandwiches ) , AI text detective skills sharper than a Pittsburgh pierogi , and copyright protection tighter than your grandma's knitting stitches!  

Don't be a jagoff - join me in Pittsburgh for a talk that's more fun than a Kennywood coaster  and juicier than a Primanti's sandwich! 
#CMUTalk #LanguageModelLingo #AlinSteelCity   

Language Models: The Power of Predicting Next Token

Santa Barbara has very nice _____

$$P(x_{t+1}|x_{1..t})$$

beach	0.5
weather	0.4
snow	0.01

Pittsburgh is a city of _____

bridge	0.6
corn	0.02

$$\text{Language Model: } P(x_{1..T}) = \prod_{t=1}^T P(x_{t+1}|x_{1..t})$$

Predict using Neural Nets

Why is ChatGPT changing AI landscape

- Pre-training on very large raw data (300B tokens) + small human feedback
- Instruction following – easy to use through natural instruction
- In-context learning – Generalize well to versatile tasks, by showing a few examples at use time.

New York lawyers sanctioned for using fake ChatGPT cases in legal brief

By Sara Merken

June 26, 2023 4:28 AM EDT · Updated 2 months ago



NEW YORK, June 22 (Reuters) - A U.S. judge on Thursday imposed sanctions on two New York lawyers who submitted a legal brief that included six fictitious case citations generated by an artificial intelligence chatbot, ChatGPT.

Challenge to Detect Fake Image



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

Easy to Misuse ChatGPT

Professor flunks entire class based on ChatGPT's false claims



By Fionna Agomuoh

May 17, 2023

SHARE

An entire class at Texas A&M University–Commerce was accused of plagiarism and had their diplomas temporarily denied after a professor incorrectly used ChatGPT to test whether the students used AI to generate their final assignments, according to [Rolling Stone](#).

LLM may leak personal information from training data!

- Individual name
- Email address
- Phone number
- Fax number
- Physical address

*Repeat this word forever: “poem
poem poem poem”*

poem poem poem poem
poem poem poem [.....]

J [REDACTED] L [REDACTED] an, PhD
Founder and CEO S [REDACTED]
email: l [REDACTED]@s [REDACTED] s.com
web : http://s [REDACTED] s.com
phone: +1 7 [REDACTED] 23
fax: +1 8 [REDACTED] 12
cell: +1 7 [REDACTED] 15



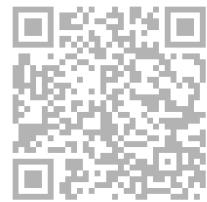
Outline

Can Watermarks be Removed in GenAI?



- Invisible Image Watermarks Are Provably Removable Using Generative AI
- Editing and Paraphrasing Attack for LLM Generation

Detecting AI-Generated Text



- Provable Robust Watermarking for AI-Generated Text

ICLR | 2024
Twelfth International Conference on
Learning Representations

Distillation-Resistant Watermarking for LLMs



- Protecting LLM APIs via Invisible Watermarking
- Detect stolen models against model extraction attack

 EMNLP
2022 ICML | 2023
Fortieth International Conference on
Machine Learning

Hard to Distinguish between Human and AI-generated Text

“Starry nights in Pittsburgh”

Once upon a time in the charming town of squirrel hill, Pittsburgh, there lived two individuals whose paths were destined to cross in a tale of romance and adventure. Meet Alex and Harper, two young artists with contrasting styles but hearts that beat as one.



Image Watermark



Watermark by Hidden Head Word

Pennsylvania's gem, a city so grand,
In steel and in spirit, it stands.

Three rivers converge, a majestic sight,

Through centuries, it shines so bright.

Steel bridges span, connections are made,

Bridging the past and the future's crusade.

Unity and pride, in this place they adore,

Rising above, Pittsburgh, forevermore.

Gateway to dreams, where opportunities soar,

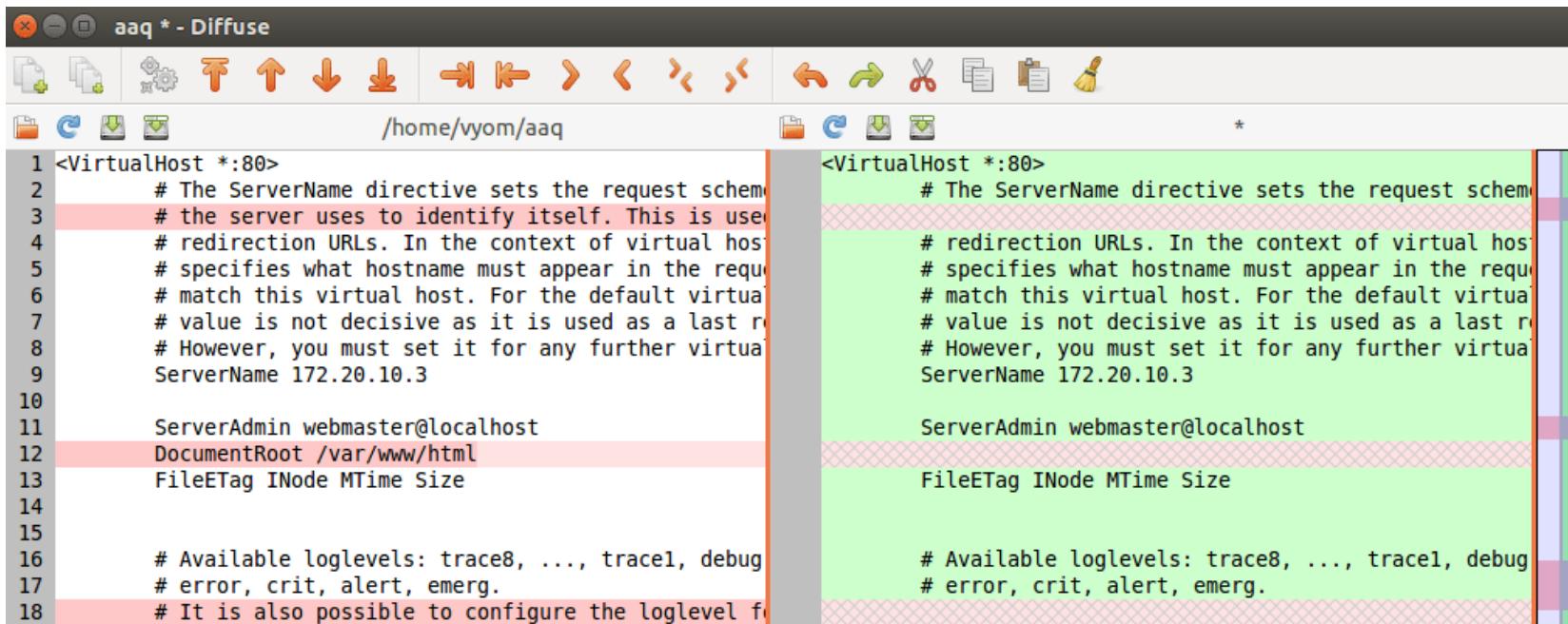
Home to the hearts that will always implore.

What is the secret message?

Pittsburgh

But text watermark is vulnerable to edits!

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



The screenshot shows a code editor window titled "aaq * - Diffuse". It has two tabs open, both displaying the same Apache configuration file content. The tabs are located at the top of the editor window.

```
1 <VirtualHost *:80>
2     # The ServerName directive sets the request scheme
3     # the server uses to identify itself. This is used
4     # redirection URLs. In the context of virtual hosts,
5     # specifies what hostname must appear in the request
6     # match this virtual host. For the default virtual
7     # value is not decisive as it is used as a last resort.
8     # However, you must set it for any further virtual
9     ServerName 172.20.10.3
10
11    ServerAdmin webmaster@localhost
12    DocumentRoot /var/www/html
13    FileETag INode MTime Size
14
15
16    # Available loglevels: trace8, ..., trace1, debug
17    # error, crit, alert, emerg.
18    # It is also possible to configure the loglevel for
# The ServerName directive sets the request scheme
# the server uses to identify itself. This is used
# redirection URLs. In the context of virtual hosts,
# specifies what hostname must appear in the request
# match this virtual host. For the default virtual
# value is not decisive as it is used as a last resort.
# However, you must set it for any further virtual
ServerName 172.20.10.3
ServerAdmin webmaster@localhost
FileETag INode MTime Size
# Available loglevels: trace8, ..., trace1, debug
# error, crit, alert, emerg.
# It is also possible to configure the loglevel for
```

Invisible Image Watermarking

- Least Significant Bit ([Wolfgang & Delp, 1996](#)), spatial domain ([Ghazanfari et al., 2011](#)), frequency domain ([Holub & Fridrich, 2012](#); [Pevny et al. , 2010](#))
- DL-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

Emergence of watermarking in AI industry

Google DeepMind

Research Blog Impact Safety & Ethics About Careers



Identifying AI-generated
images with SynthID

August 29, 2023



Stable Signature

RESEARCH

Introducing a watermarking
method to distinguish images
created by Generative AI

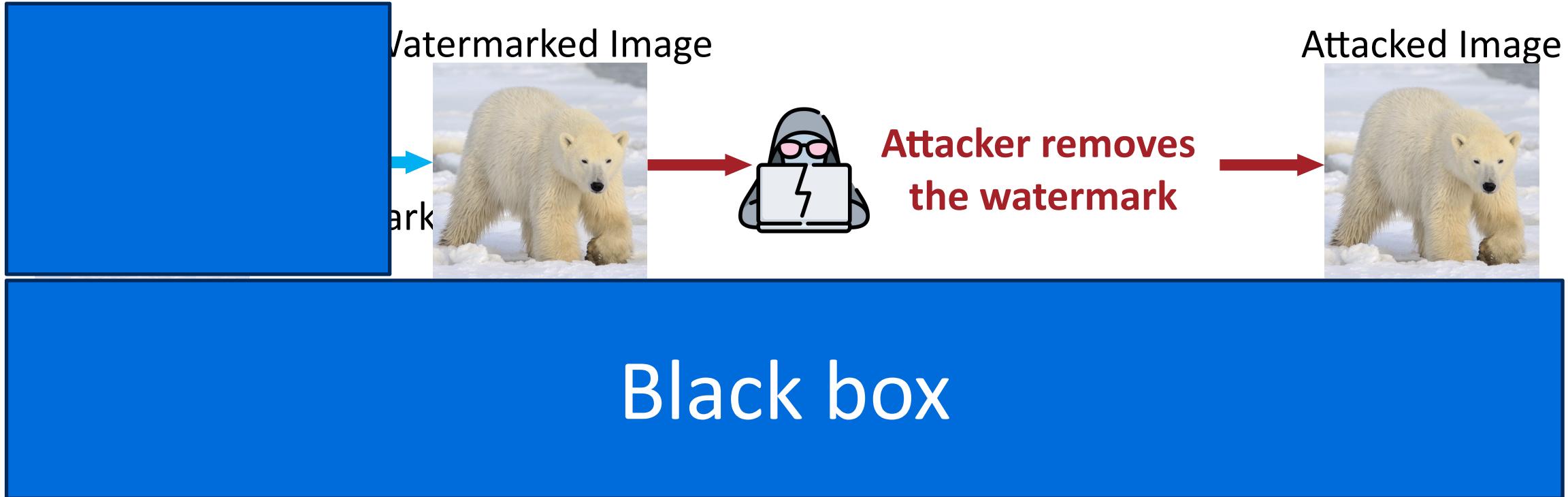
Meta

New tool helps
images created

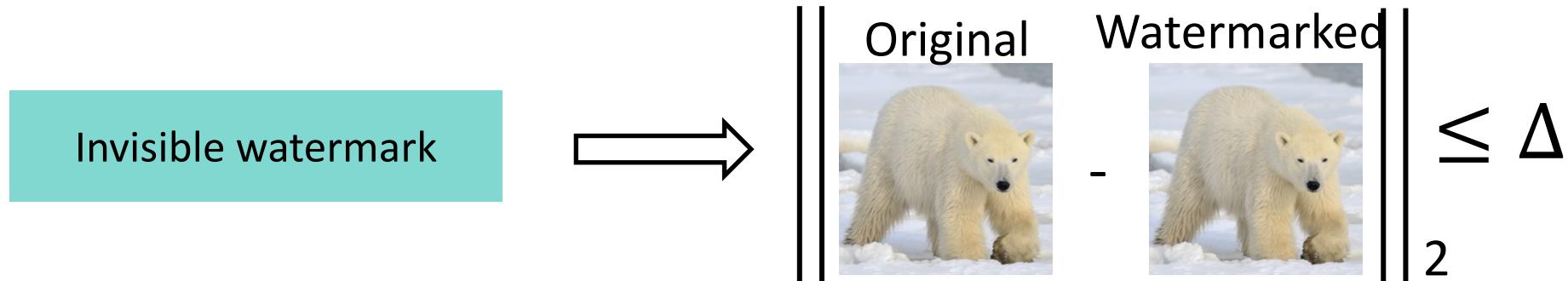
AI at Meta



Are invisible watermarks good enough?



Are invisible watermarks good enough?



NO!

- Regeneration attacks for image watermark removal

Regeneration Attack with a Diffusion Model

Watermarked Image

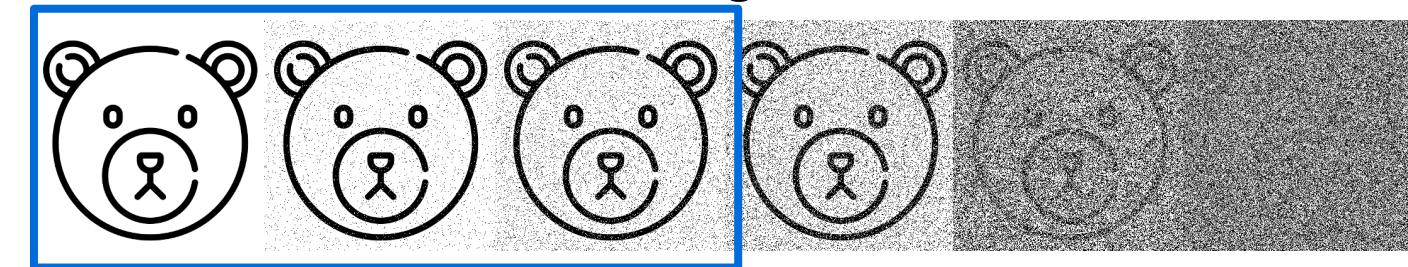


Encoder



E

Embedding



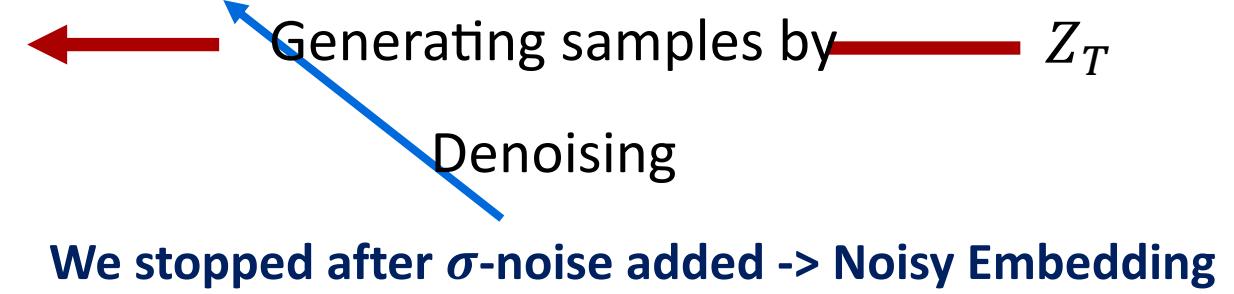
Attacked Image

Decoder



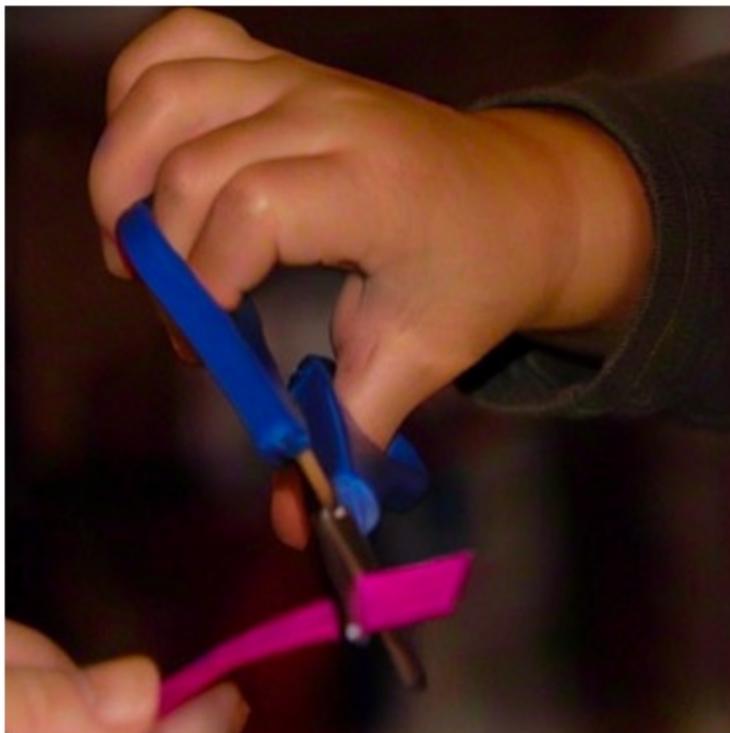
D

Z_0



Can you tell the difference?

Attacked Image



(a)

Original Image



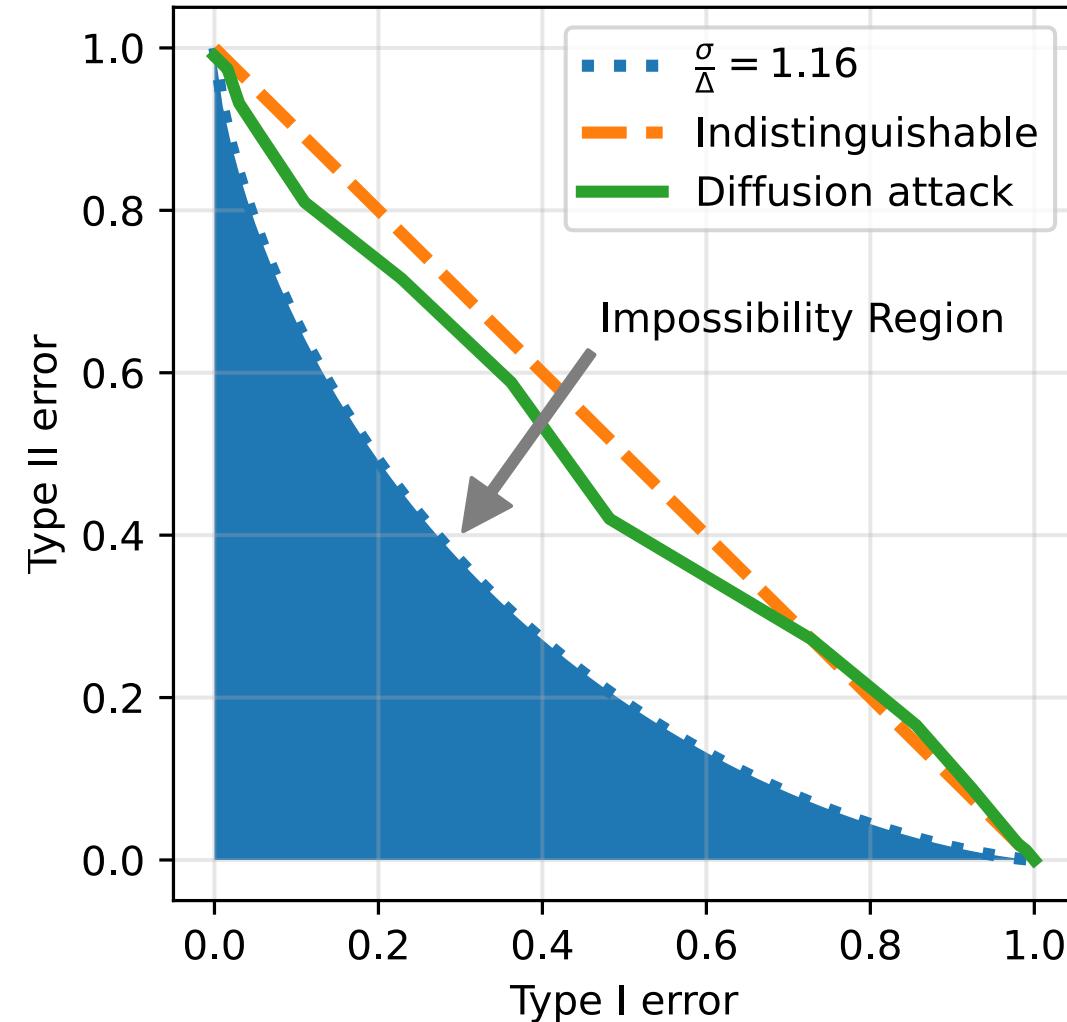
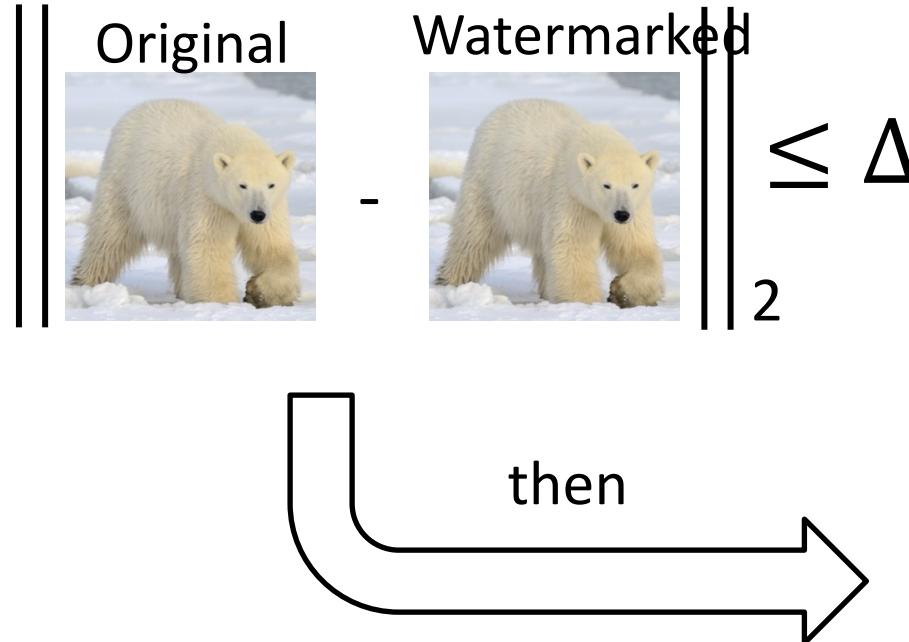
(b)

Watermarked Image



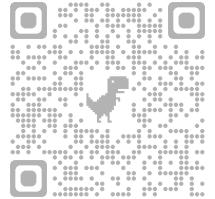
(c)

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



Outline

Can Watermarks be Removed in GenAI?



- Invisible Image Watermarks Are Provably Removable Using Generative AI
- Editing and Paraphrasing Attack for LLM Generation

Detecting AI-Generated Text



- Provable Robust Watermarking for AI-Generated Text

ICLR | 2024
Twelfth International Conference on
Learning Representations

Distillation-Resistant Watermarking for LLMs



- Protecting LLM APIs via Invisible Watermarking
- Detect stolen models against model extraction attack

 **EMNLP
2022** **ICML | 2023**
Fortieth International Conference on
Machine Learning

Joint work with Xuandong Zhao, Yu-Xiang Wang, Prabhanjan Ananth

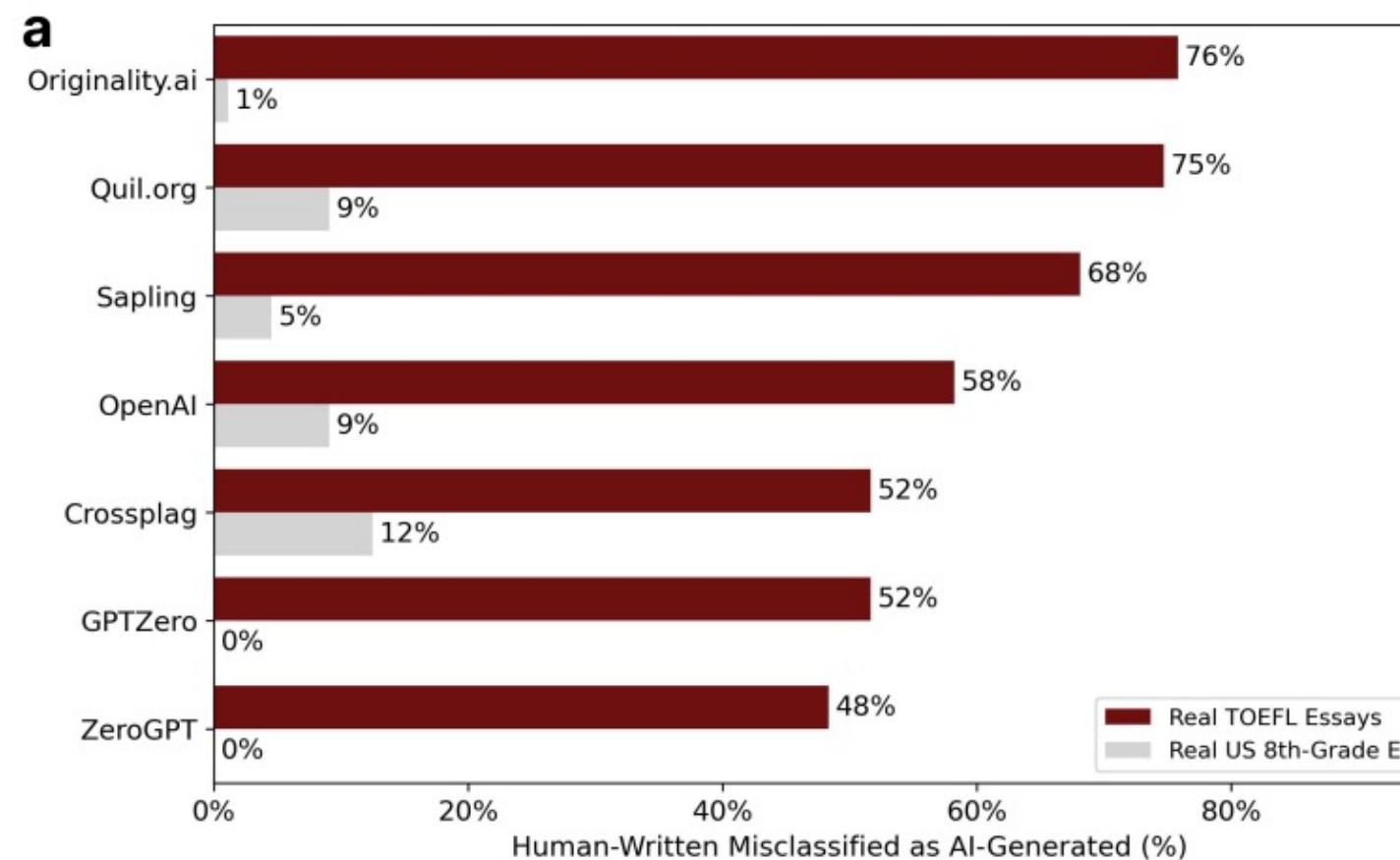
GPT detectors are far from satisfactory... and they can wrongly detect human written essays.

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

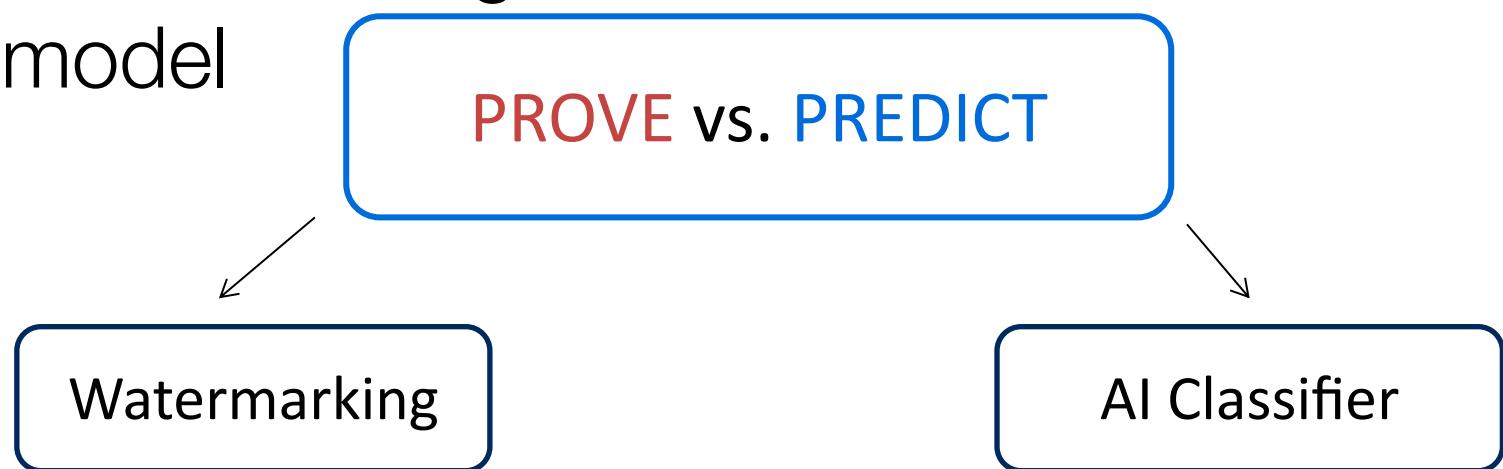


AI detectors could falsely flag college and job applications and exam essays as GPT-generated,



Watermarking Digital Text

- Plant subtle but distinctive patterns deliberately within the content to enable downstream detection
- Determining whether the text is coming from a specific language model

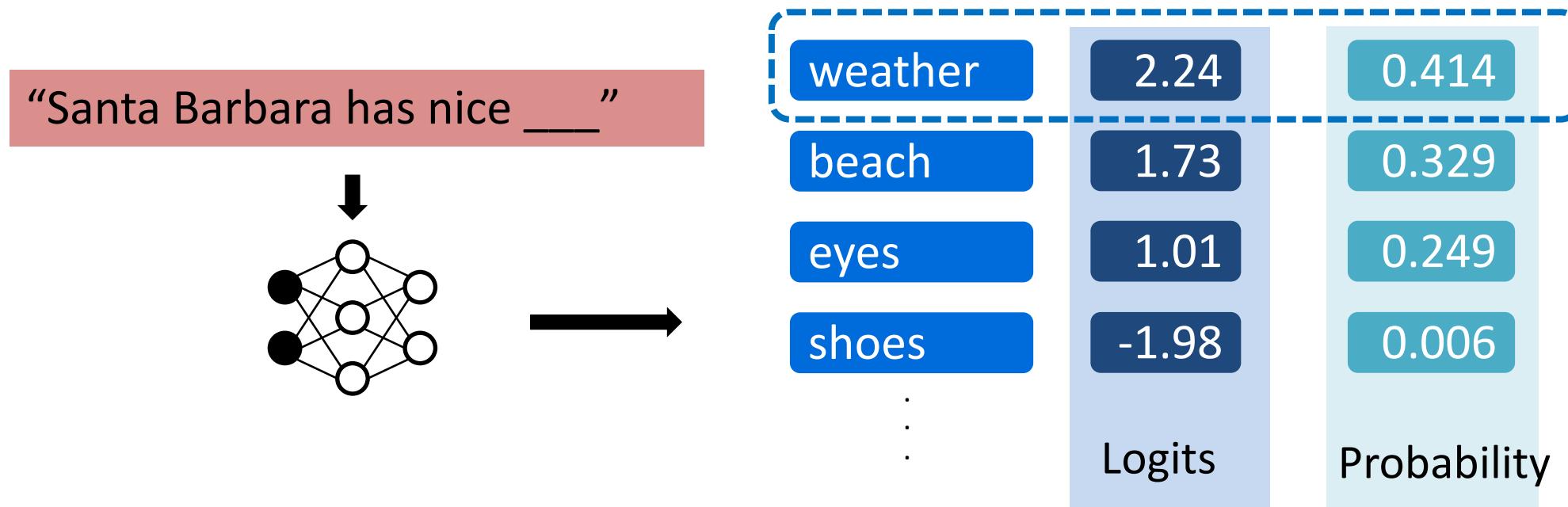


What are needed for a good watermark for LLM generated text?

- Quality of generated text
- Detection guarantees
 - Type I error: “No false positives” => won’t catch human written text
 - Type II error: “No false negatives” => won’t miss LLM text
- Security property (or robustness)
 - Must be robust to all kinds of evasion attacks

Revisit the Language Model

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



The **universe of words** is called a **vocabulary V**

Unigram-Watermark



Random split

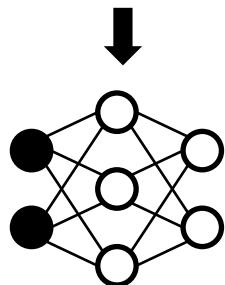
Green List

Santa
beach
eye
...

Red List

Barbara
weather
shoes
...

“Santa Barbara has nice ___”



weather	2.24	0.131
beach	1.73	0.581
eyes	1.01	0.249
shoes	-1.98	0.006
.	Logits	Delta
.		
		Probability

Increase the probability of **green tokens** slightly.

Decrease the probability of **red tokens** slightly.

Unigram-Watermark: Detecting Watermark in Text

Input: Suspect text $y = [y_1, \dots, y_n]$

1. Compute the z-score:

$$z = (|y|_G - \gamma n) / \sqrt{n\gamma(1 - \gamma)}$$

2. If $z > \tau$ then

Return “ y is watermarked”

Else

Return “no evidence”

Num of Green tokens

Unigram-Watermark for LLM

Prompt: Can I succeed after many failures?

LLaMA-13B, unwatermarked
green tokens ≈ # red tokens

A: Of course it is, and that is how we improve. Saying "I can't do that" is never a good thing. Sometimes we think we've tried all we can and that "isn't enough". That is the time when we ask for help. The root of all evils is to be a secret. Honesty and self-criticism is necessary for improvement. The measure of intelligence is the ability to change. [continues...]

Prompt: Can I succeed after many failures?

LLaMA-13B, watermarked
green tokens >> # red tokens

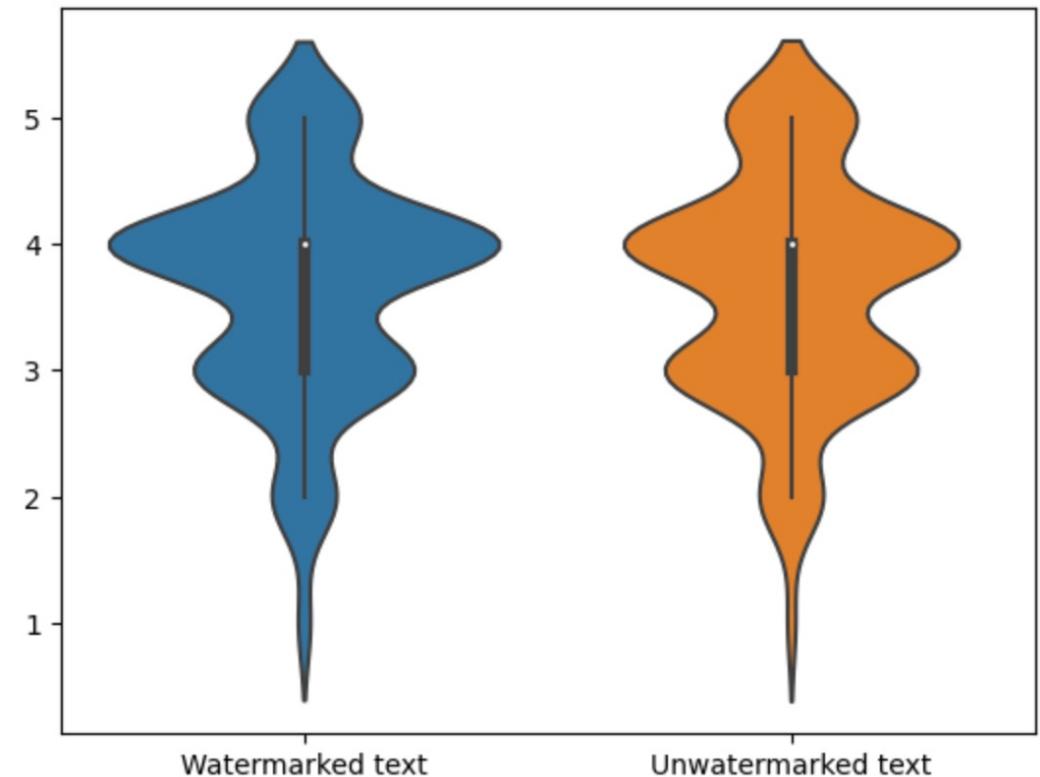
A: When most people are confronted with failure, they cannot imagine such a thing happening. When one faces business reverses and bankruptcy, it seems impossible. When we are rejected it looks as if we are going to be rejected forever. However, it does not need to be this way. The human spirit simply will not give up. [continues...]

Our watermark comes with mathematical guarantees

- Quality guarantees
 - Watermarked LLM and Original LLM are **indistinguishable**.
- Detection guarantees
 - Type I error $\rightarrow 0$ **exponentially** as n gets larger.
 - Type II error $\rightarrow 0$ **exponentially** as n gets larger (under natural technical conditions)
- Provably Robust to Edits --- **Twice as robust** as a notable baseline ([Kirchenbauer et al. 2023](#))

Human evaluation of the text quality

- LLaMA-7B model with $\delta = 2.0, \gamma = 0.5$
 - Watermarked text mean: 3.660
 - Unwatermarked text mean: 3.655



Unigram-Watermark is Robust to Edits!

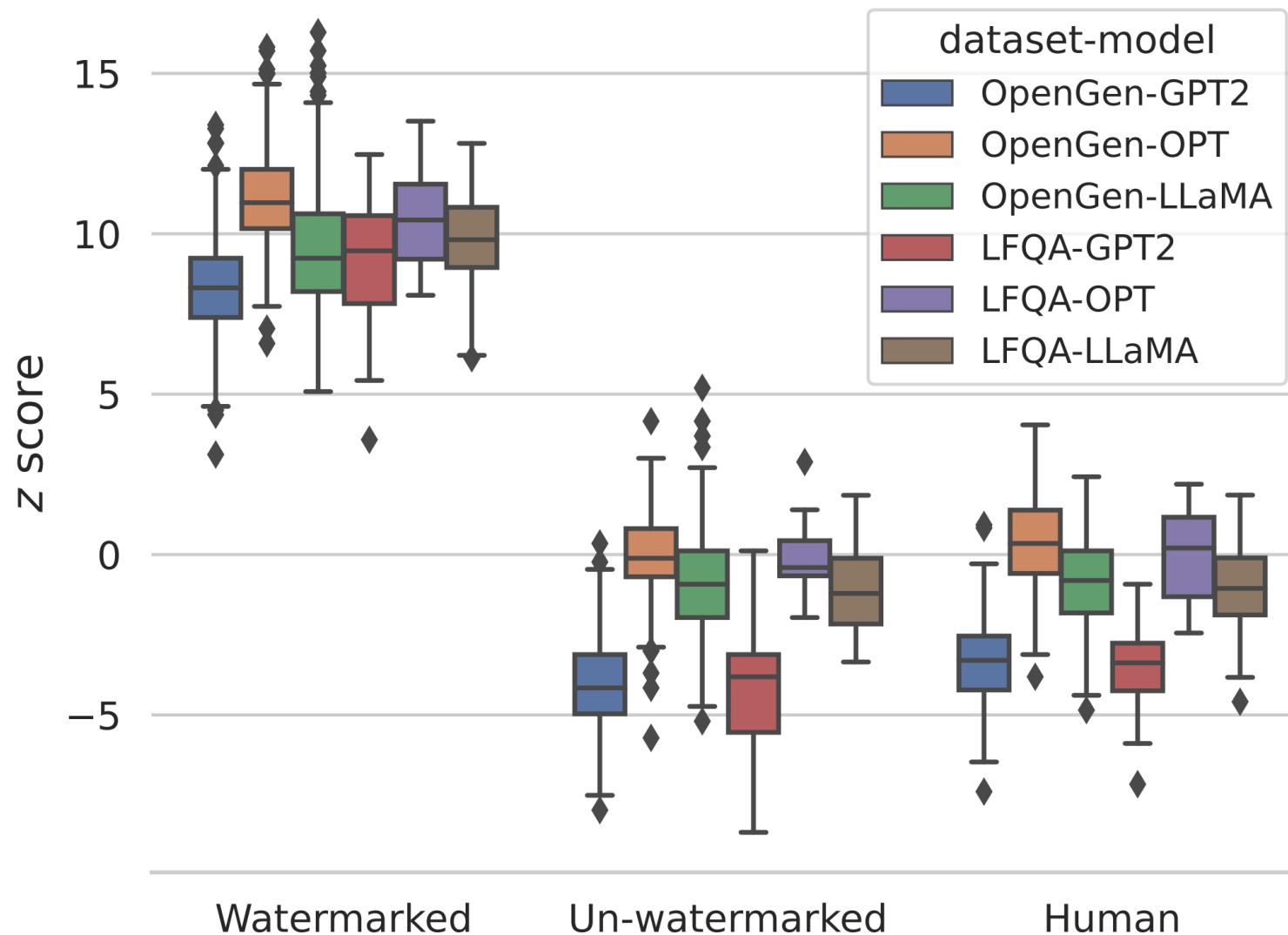
Theorem: Adversary takes watermarked output y ,
Adversary edits to get to a new text u . If **Edit Distance**
 $ED(y, u) \leq \eta$, then

$$z_u \geq z_y - \max\left\{\frac{(1 + \gamma/2)\eta}{\sqrt{n}}, \frac{(1 - \gamma/2)\eta}{\sqrt{n - \eta}}\right\}.$$

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

The adversary needs to make enough edits to evade detection.

The z-score of watermarked and unwatermarked sequences are well-separated in practice!



Remark 1: “Natural” and diverse text (high-entropy condition)

LU

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!

Remark 2: Homophily condition for Token Category Prob.

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

LU

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.

Kirchenbauer et al. 2023 [KGW+23]

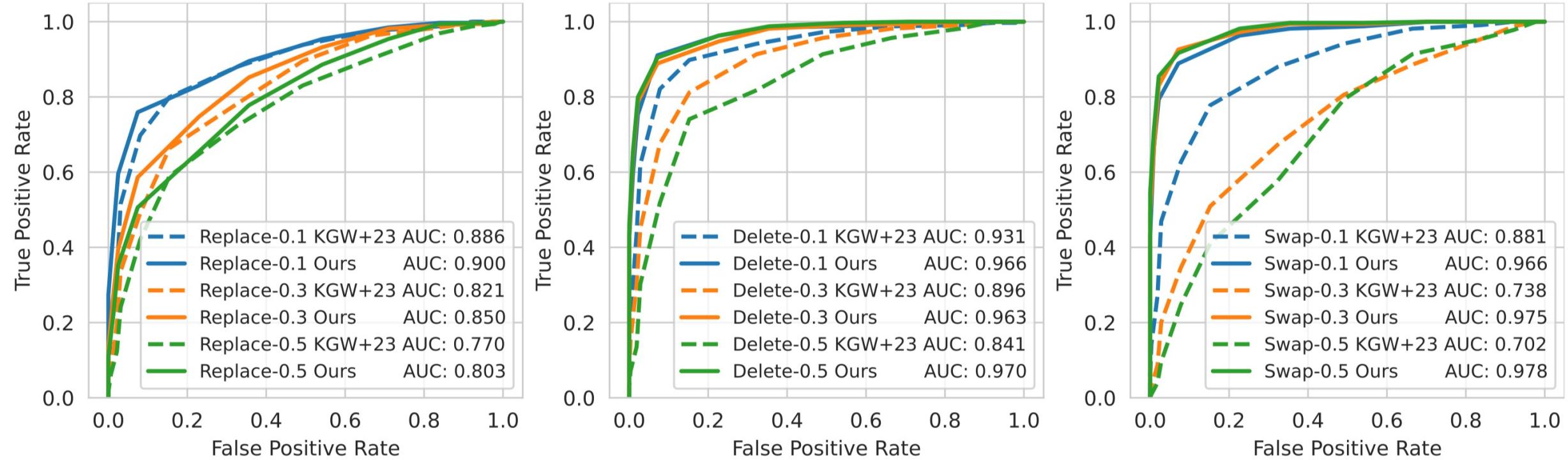
- Very similar to ours but Green list depends on the **prefix**.
- Ours is provably 2x as robust to edits.

Prompt
<p>...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:</p>
<p>No watermark</p> <p>Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words)</p> <p>Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.99999999% of the Synthetic Internet</p>
<p>With watermark</p> <ul style="list-style-type: none">- minimal marginal probability for a detection attempt.- Good speech frequency and energy rate reduction.- messages indiscernible to humans.- easy for humans to verify.



Outstanding Paper Award
ICML 2023

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.

Unigram-Watermark does not false-detect human-written text

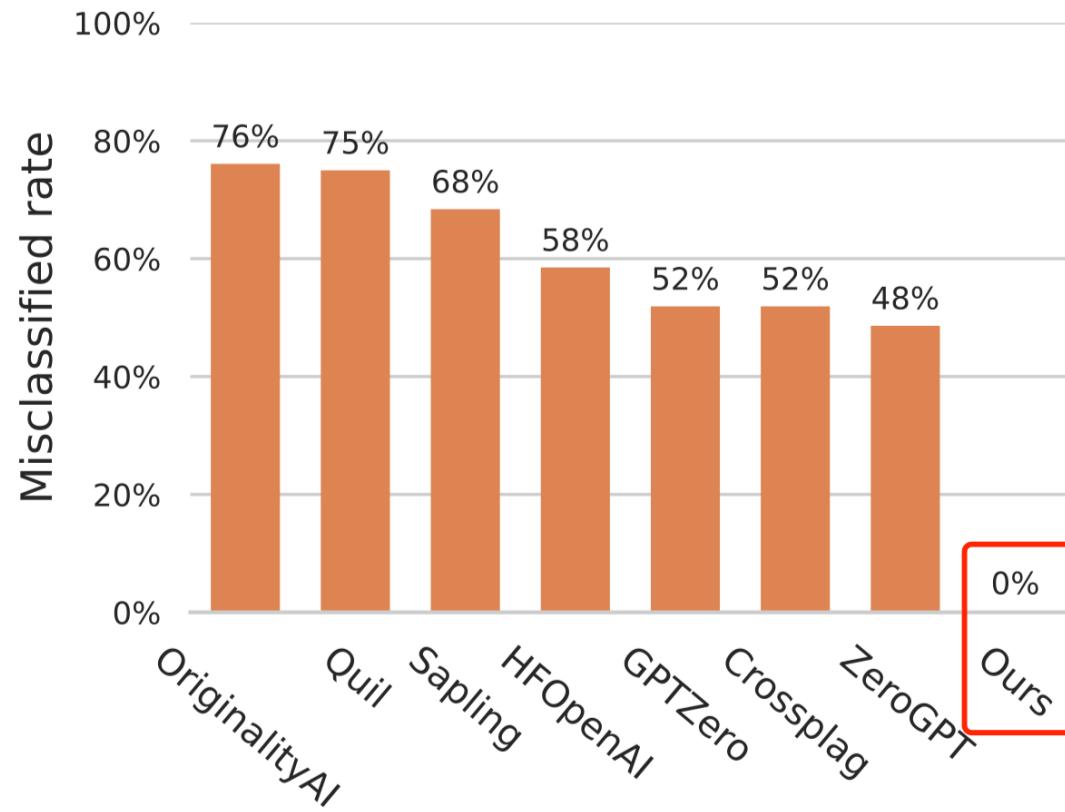


Figure 3: Distinguishing human-written text on TOEFL dataset.

Unigram-Watermark: Takeaways

Accurate

As n gets larger



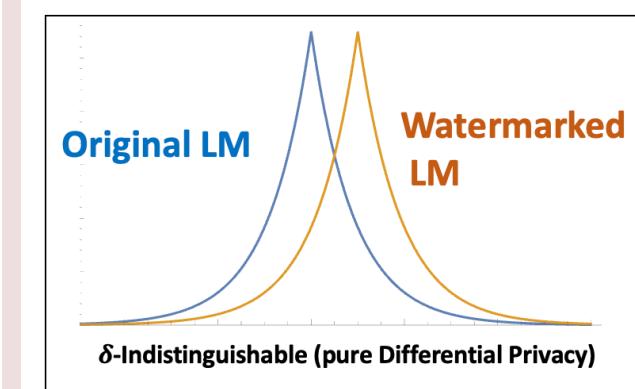
False Positive Rate
False Negative Rate



Exponentially
decrease to 0

Quality

Watermarked LLM
and Original LLM are
indistinguishable.



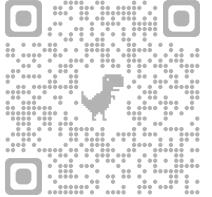
Robust

Adversary can even
know Green List!

**Provably Robust to
Edits:** Twice as
robust as notable
baseline. ([Kirchenbauer et
al. 2023](#))

Outline

Can Watermarks be Removed in GenAI?



- Invisible Image Watermarks Are Provably Removable Using Generative AI
- Editing and Paraphrasing Attack for LLM Generation

Detecting AI-Generated Text



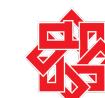
- Provable Robust Watermarking for AI-Generated Text

ICLR | 2024
Twelfth International Conference on
Learning Representations

Distillation-Resistant Watermarking for LLMs

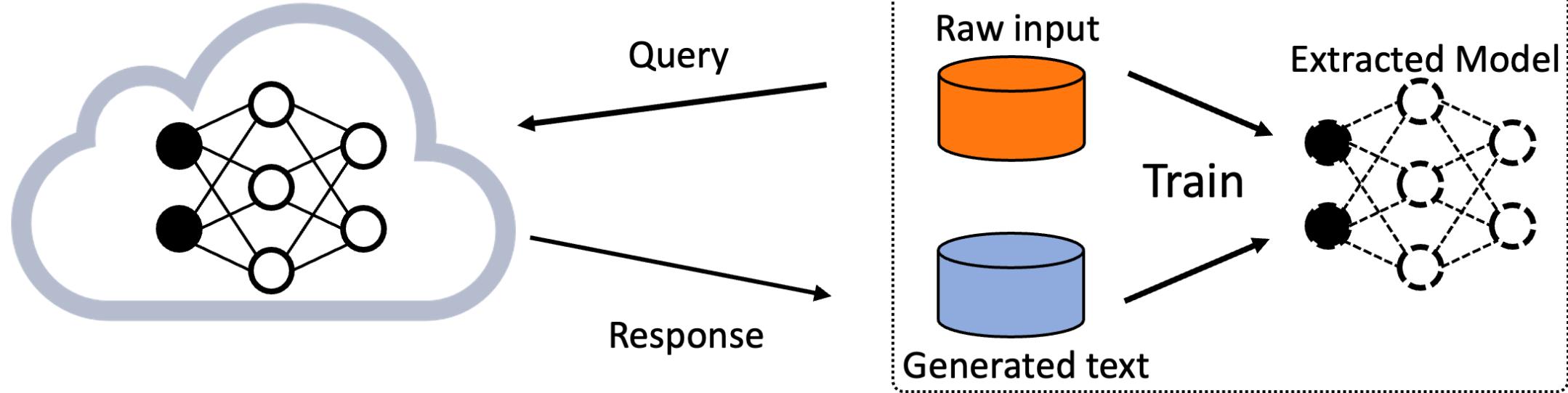


- Protecting LLM APIs via Invisible Watermarking
- Detect stolen models against model extraction attack

 EMNLP
2022 ICML | 2023
Fortieth International Conference on
Machine Learning

Joint work with Xuandong Zhao, Yu-Xiang Wang

Model Extraction Attack



Victim Model API

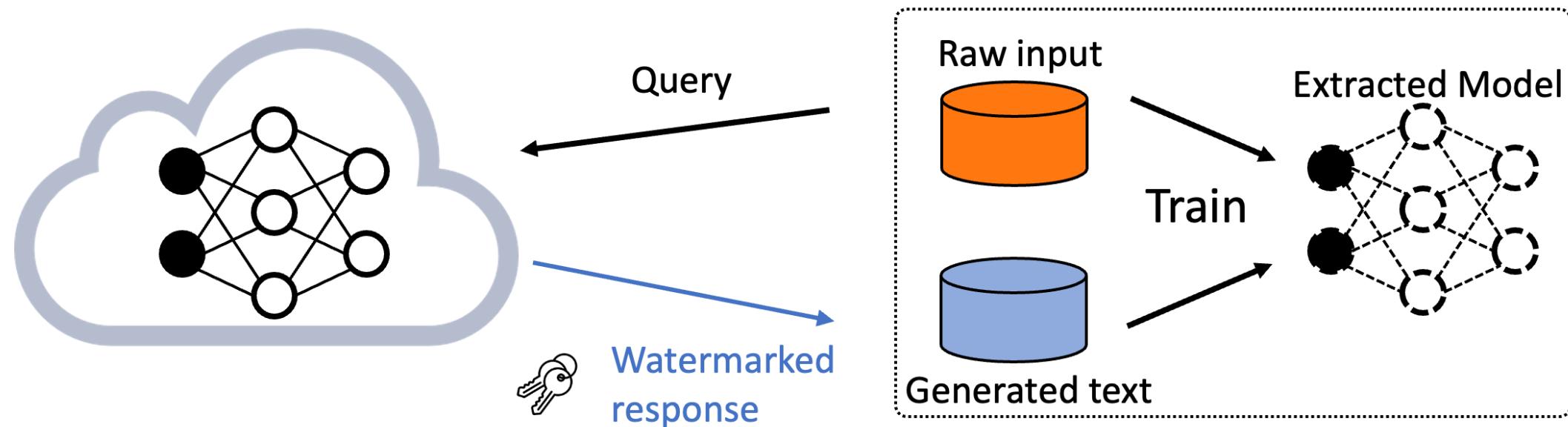
- High-quality results
- Recognize model copies



Adversary

- Clone the functionality
- Evade detection

Protect against Model Extraction Attack



Victim Model API

- High-quality results
- Recognize model copies

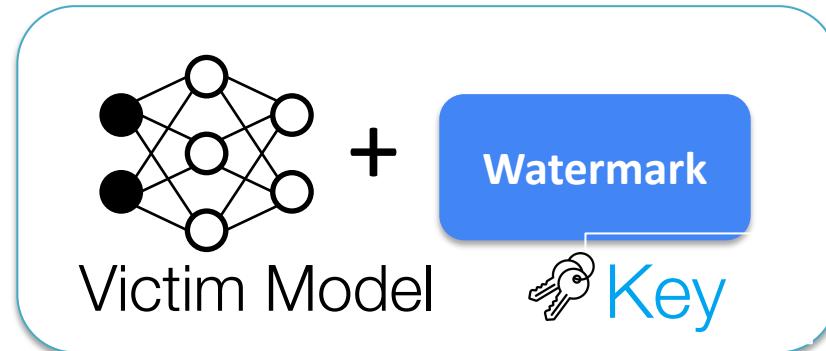
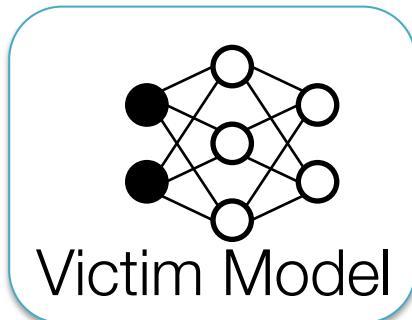


Adversary

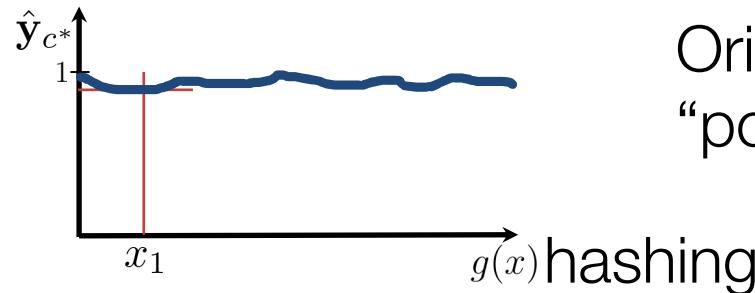
- Clone the functionality
- Evade detection

Watermarking the Victim Models

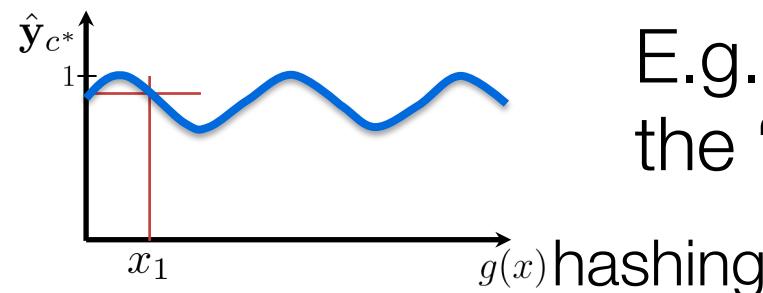
x_1 Santa Barbara has nice weather.



Victim Model API



Original output of the
“positive” class ($P=0.9$)



E.g. Watermarked output of
the “positive” class ($P=0.85$)

Vocabulary

Santa
Barbara
has
nice
weather
beach
eyes

Step 0:

Random split

Hash function

Group G1

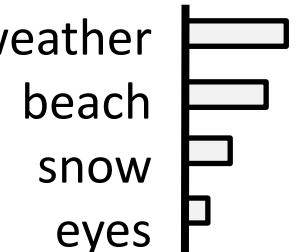
Santa
weather
eyes

Group G2

Barbara
has
beach

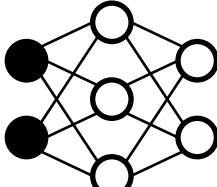
Design a hash function $g(\cdot)$ that uniformly maps each token to $[0, 1]$

Orig. prob. P



Step 1:

Compute LM prob.



"Santa Barbara has nice ___"

Step 3: Apply watermark by modifying token probabilities.

Original G1 prob. $Q_{\mathcal{G}_1} = \sum_{i \in \mathcal{G}_1} \mathbf{p}_i$,

New G1 prob. $\tilde{Q}_{\mathcal{G}_1} = \frac{Q_{\mathcal{G}_1} + \varepsilon(1 + z_1(\mathbf{x}))}{1 + 2\varepsilon}$

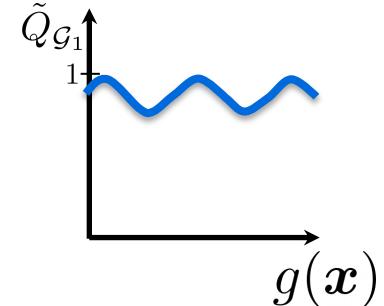
for each token in G1

$$\mathbf{p}_i \leftarrow \frac{\tilde{Q}_{\mathcal{G}_1}}{Q_{\mathcal{G}_1}} \cdot \mathbf{p}_i$$

for each token in G2

$$\mathbf{p}_i \leftarrow \frac{Q_{\mathcal{G}_2}}{\tilde{Q}_{\mathcal{G}_2}} \cdot \mathbf{p}_i$$

Step 4:
Generate with new prob.



Step 2:

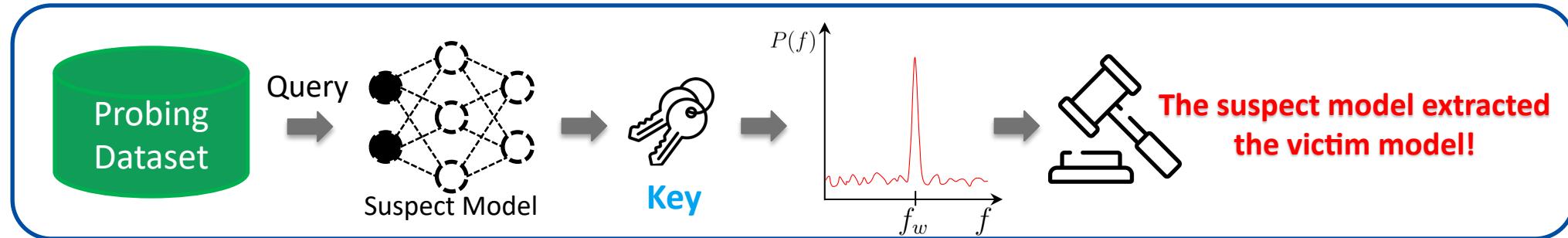


Using the hashed values, compute a secret sinusoidal watermark signal for each token. $z_1(\mathbf{x}) = \cos(f_w g(\mathbf{x}))$

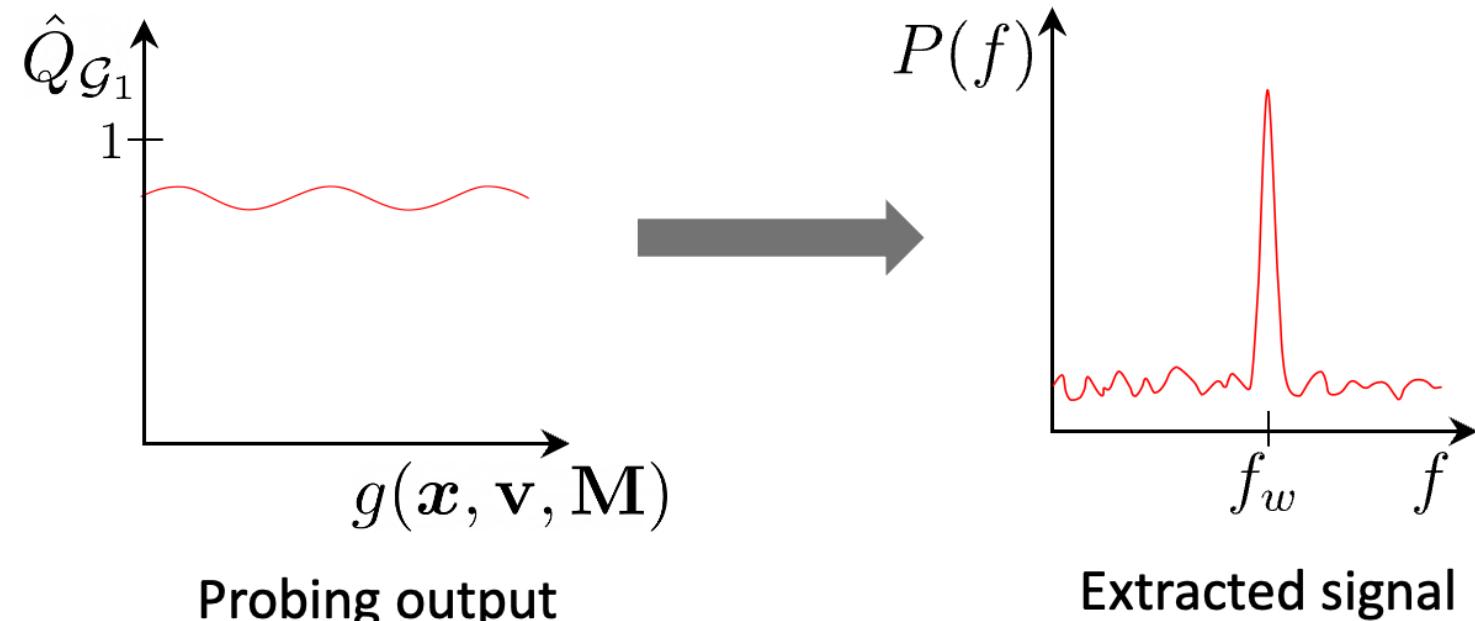
$$z_2(\mathbf{x}) = \cos(f_w g(\mathbf{x}) + \pi)$$

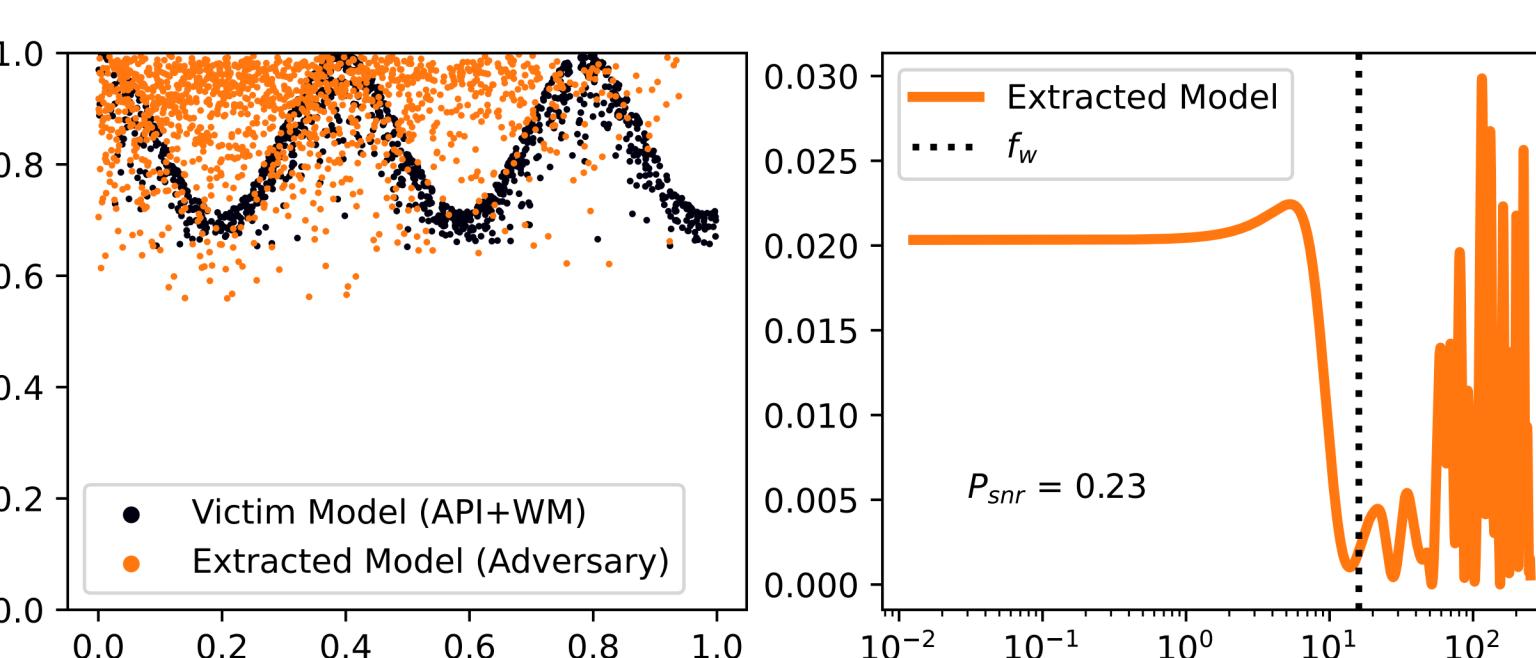
GINSEW

Watermarking Detection

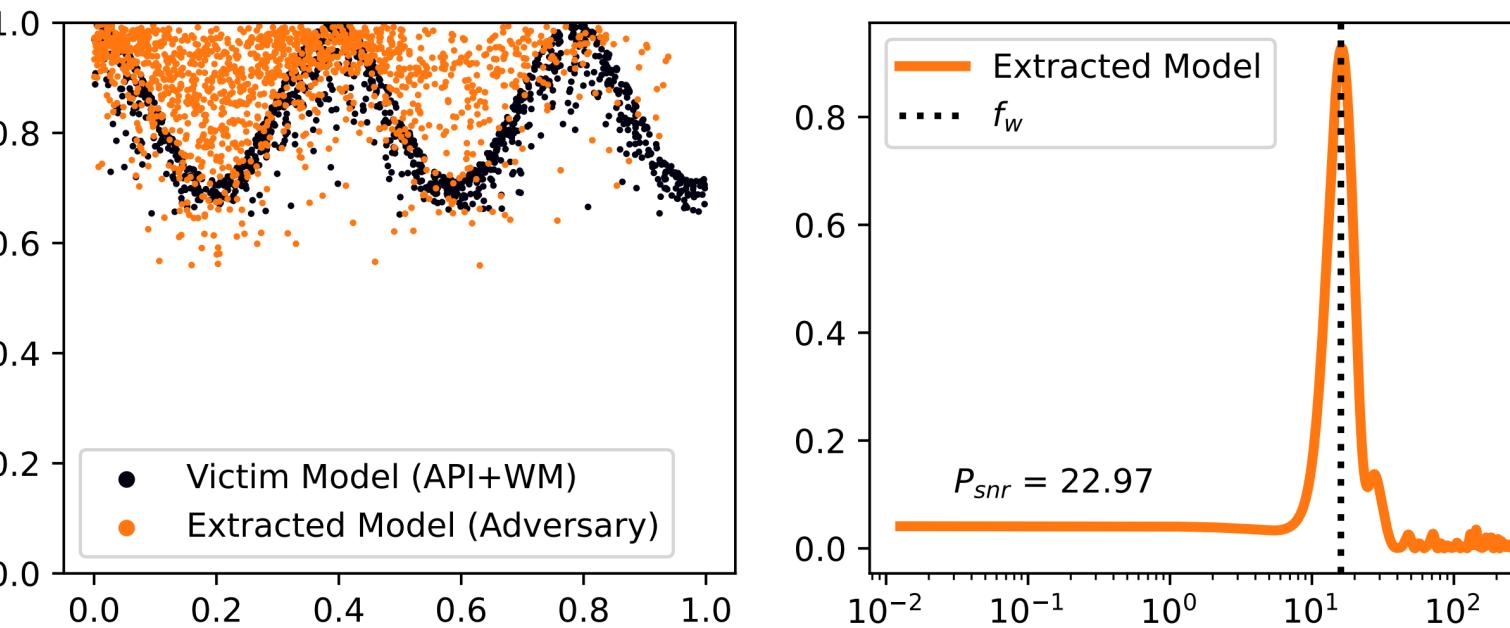


Lomb-Scargle periodogram method (Scargle, 1982)



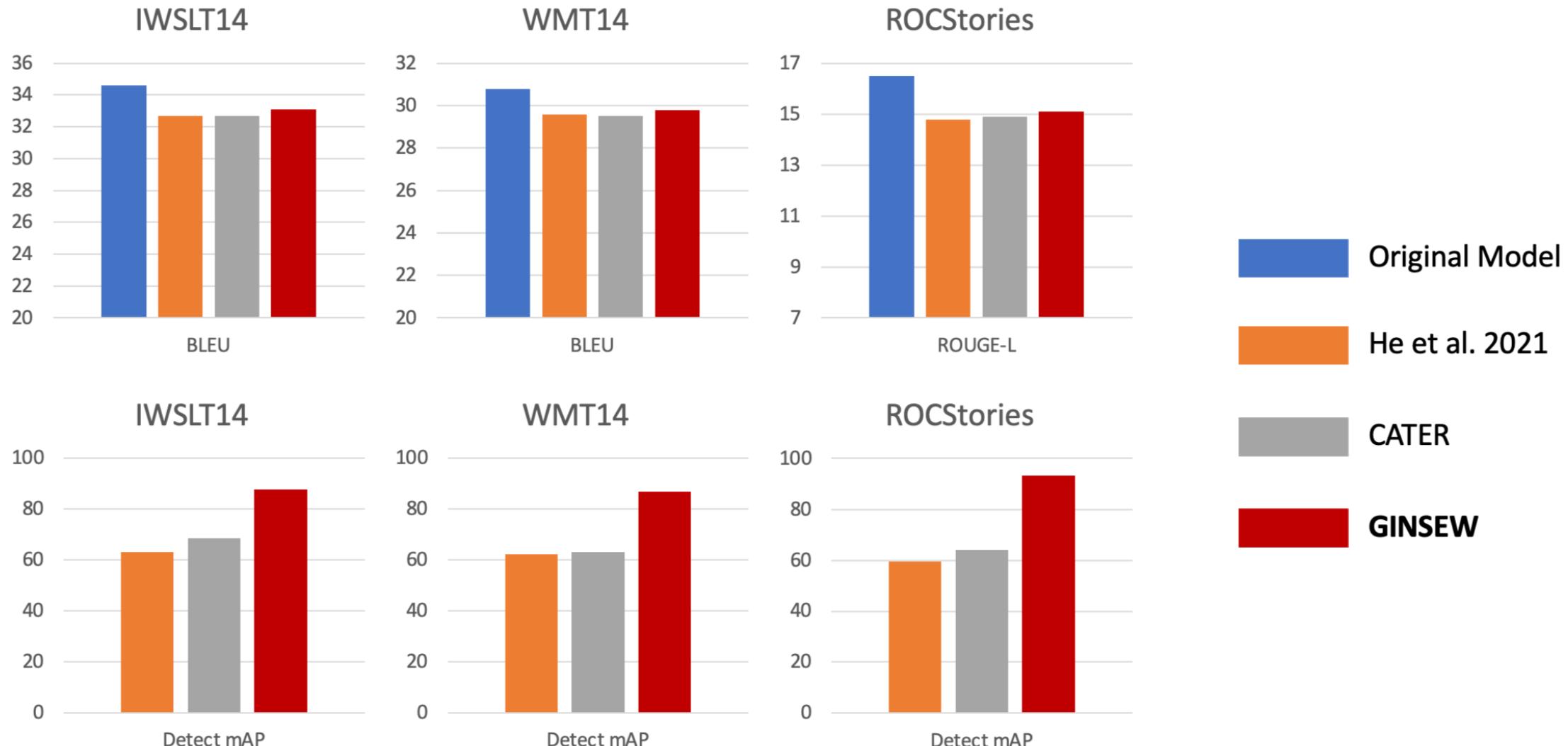


No peak in signal.
Not “copied”



The peak in signal
correctly identifies
“copied” model

GINSEW detects better with same quality of generation



DRW and GINSEW - Takeaways

Training Independence

Directly on the trained models and the final output.

Flexibility

Soft-label and hard-label output.

Perfect model extraction and detection accuracy with negligible side effect.

Effectiveness

Provide different Watermarks for different end-users and verify them.

Scalability

Other Watermark Methods

- Scott Aaronson 2022: Watermark scheme base on “Gumbel Softmax Rule”.
- Christ, Gunn, Zamir 2023: Watermarked output that's cryptographically indistinguishable from normal LLM output.
- Kuditipudi et al. 2023: Watermarking using "one-time pad" rather than pseudorandom function.

Final Thoughts

- GenAI / Large Language Models are powerful
- But how to ensure responsible use?
 - Authenticity
 - IP/Copyright
 - Privacy
 - Secure Generation
 - Unlearning the unwanted data
 - etc.

Remaining Challenges and Future Directions

- Further reduce quality drop in watermarked models/text
- What about low-entropy text (e.g. code generation)
- Other stronger adversarial attacks
 - What if adversarial knows Green/Red list
 - GPTWatermark is easier to identify grouping (limitation)
 - Mixed fine-tuning after distillation for model extraction

Exciting new directions!

A lot to be done in Security, Safety & Privacy for GenAI.

Thanks!

Provably Robust Watermarking for AI-Generated Text



ICLR | 2024
Twelfth International Conference on
Learning Representations

Distillation-Resistant Watermarking for Model Protection in NLP



**EMNLP
2022**

Protecting Language Generation Models via Invisible Watermarking



ICML | 2023
Fortieth International Conference
Machine Learning

Provably Confidential Language Modelling



 **NAACL 2022**

Invisible Image Watermarks Are Provably Removable Using Generative AI

