NEURAL INFORMATION PROCESSING SYSTEMS

Website; Q&A

Watermarking for Large Language Models

Part I: Introduction

Xuandong Zhao
UC Berkeley

Yu-Xiang Wang
UC San Diego

Lei Li
CMU

Large Language Models

2020

2023

2024

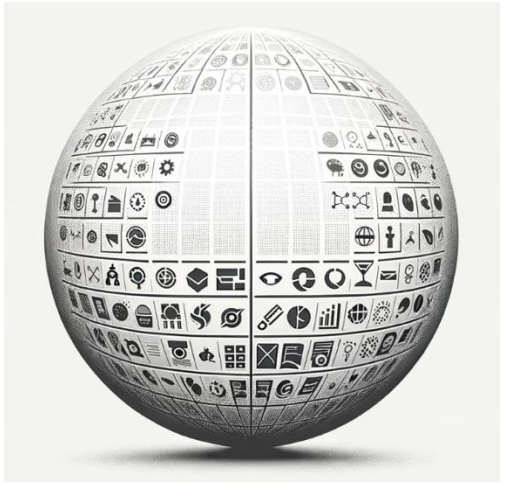
Google Gemini

Supercharge your creativity and productivity

Chat to start writing, planning, learning and more with Google AI

Translate

Chat

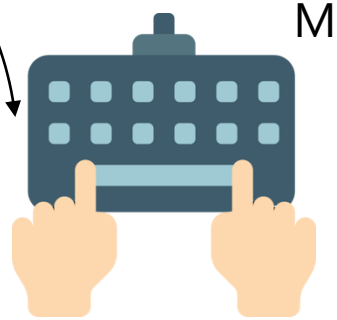


Summarize

Math Calculation

Polish Email

Large Language Models



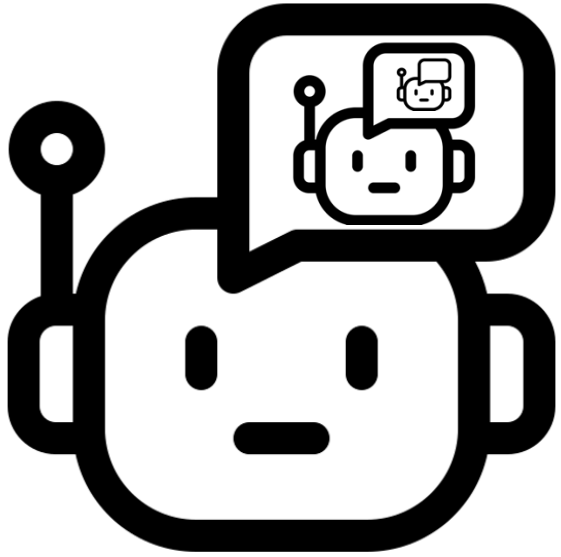
Write Code

Risks of LLMs

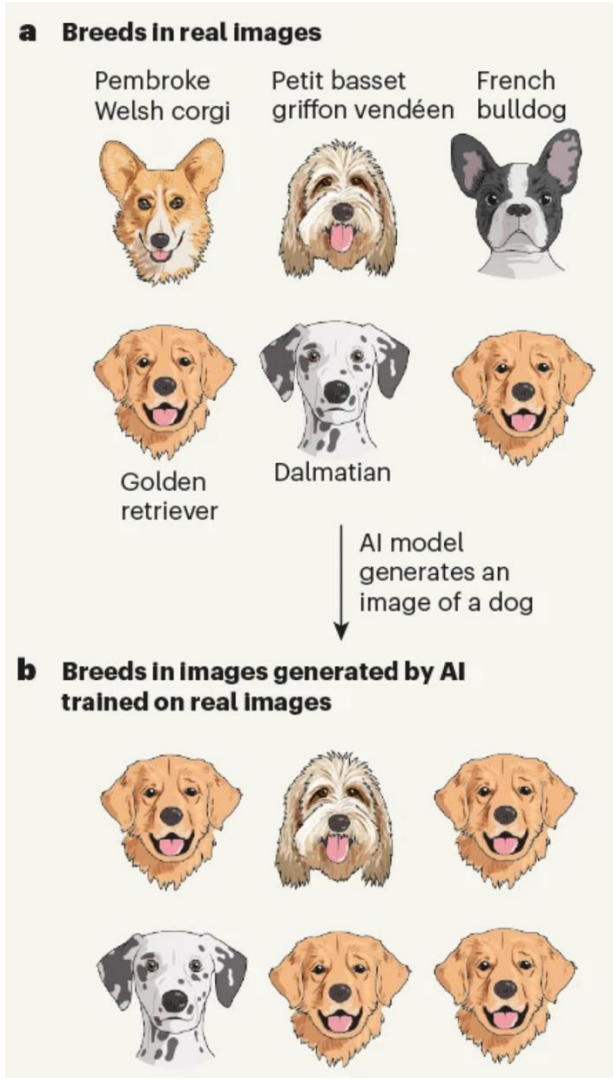
- Fake news…
- Bogus case law…
- Malware…
- Scams…
- Plagiarism…
- Private data leaks…
- …

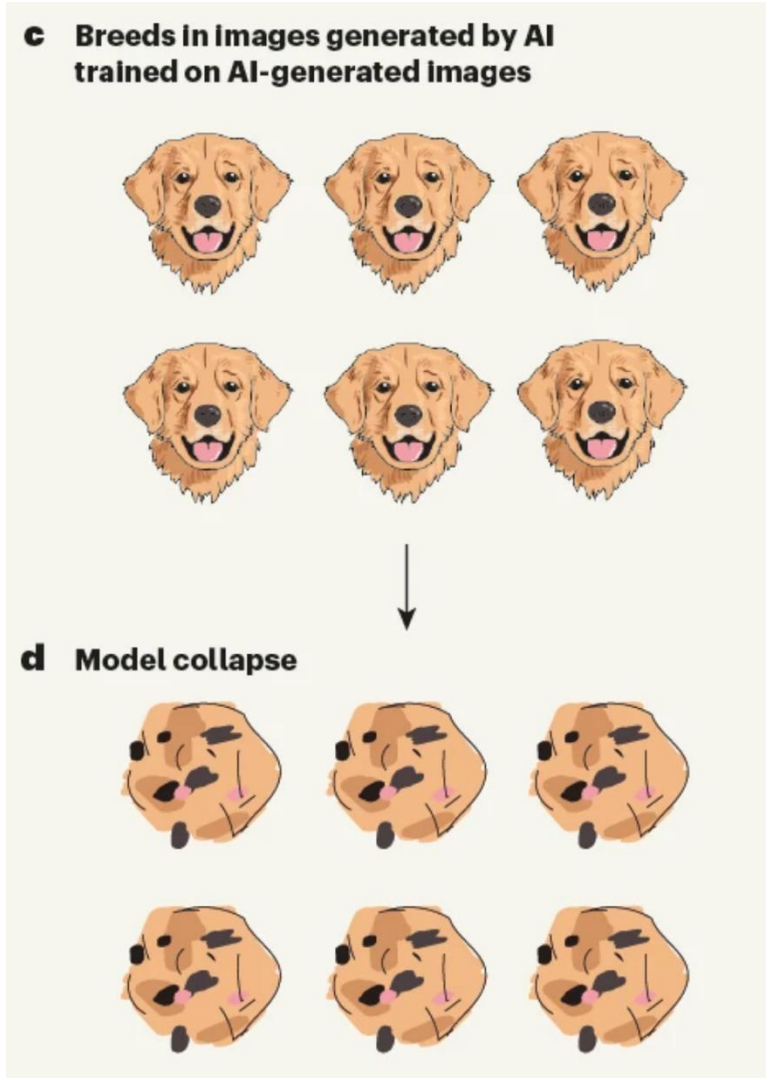
Why do we need to detect AI-generated text?

Why do we need to detect AI-generated text?



Model Degeneration
or
Model Collapse





Shumailov et al. AI models collapse when trained on recursively generated data. Nature 2024

Why do we need to detect AI-generated text?

Can you distinguish human vs. machine generated text?

Through the town, and past the lights,
Oh, how the bells do ring!
They chime with glee
For you and me
As carols we joyfully sing.

Over the river, and through the wood,
Oh, how the wind does blow!
It stings the toes
And bites the nose
As over the ground we go.

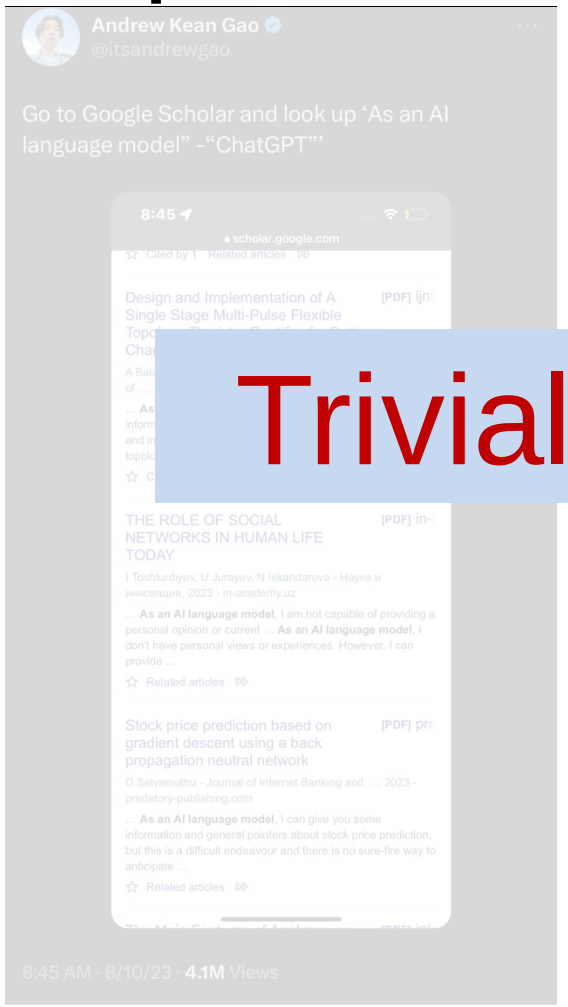

Machine

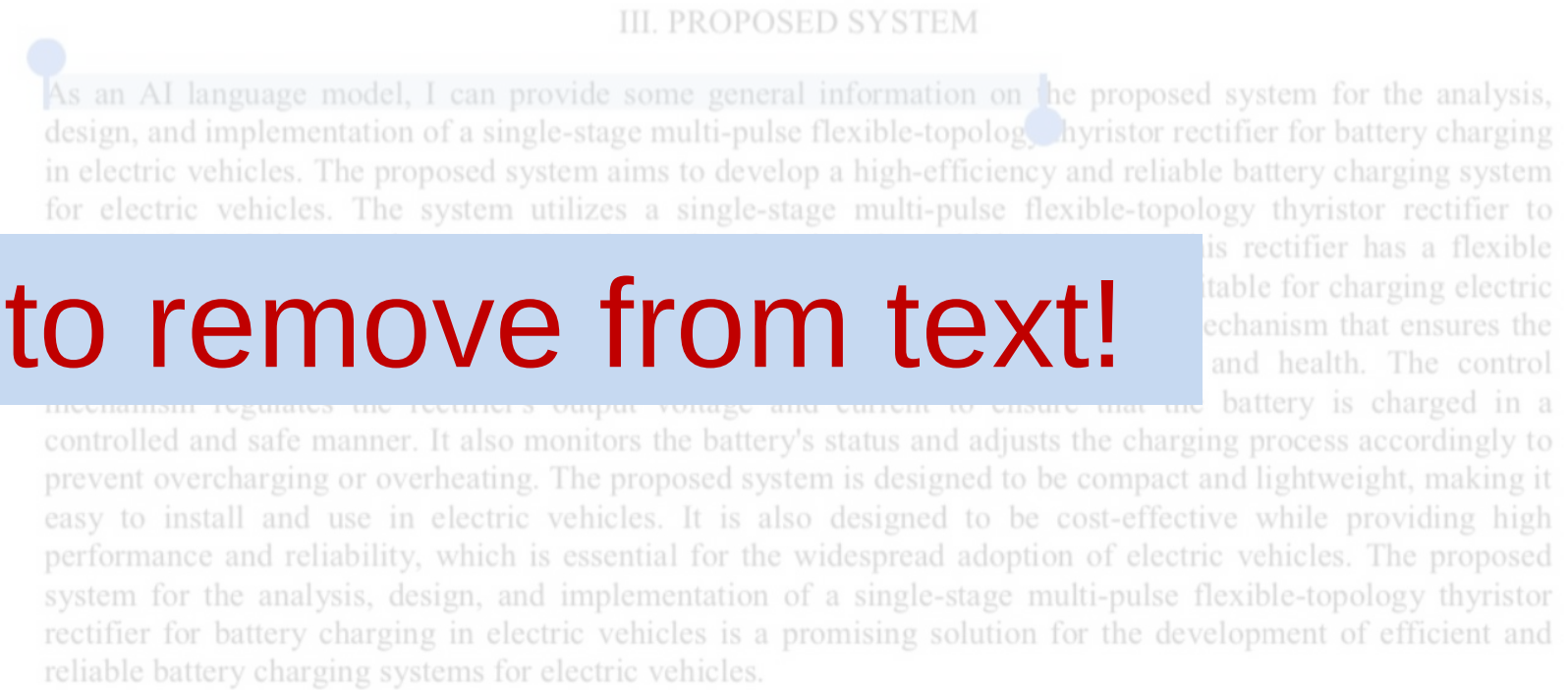

Human

Child, Lydia Maria. “Thanksgiving Day.” 1844.

How to detect AI-generated text?

- Add prefix: “As a large language model…”





Trivial to remove from text!

https://x.com/itsandrewgao/status/1689634145717379074

How to detect AI-generated text?

- Maintain a database of all completions

Expensive? Privacy?
Open-source models?

How to detect AI-generated text?

- **Train classification models** [GPTZero, Turnitin, …]

Too many false positives?
Out-of-distribution data?
AI development?

Watermarking is a promising solution!

Plant subtle but distinctive signals deliberately within the content to enable downstream detection

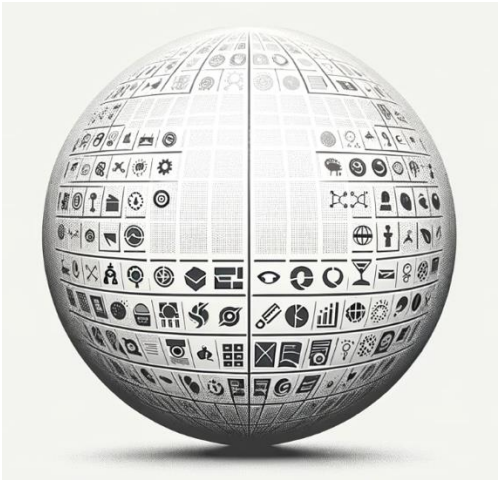
Part II: Text Watermarking

Watermarking vs. AI Classifier

Active

Passive

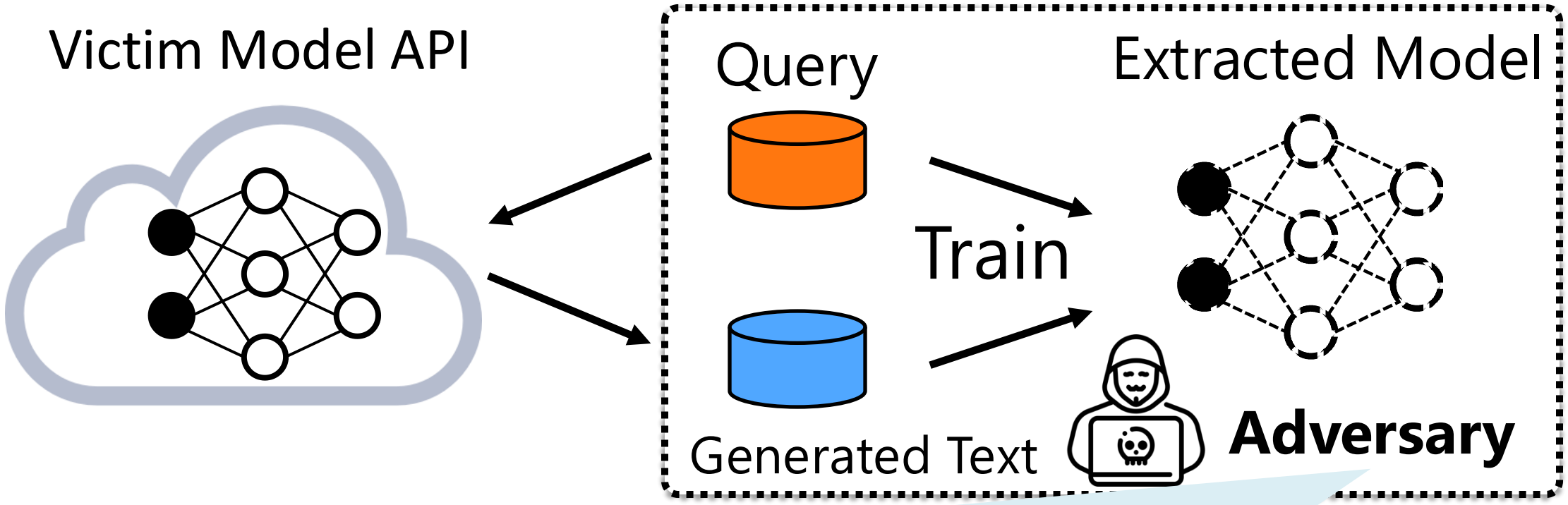
Intellectual Property of LLM



I want to steal the model!

Model Stealing/Extraction Attack

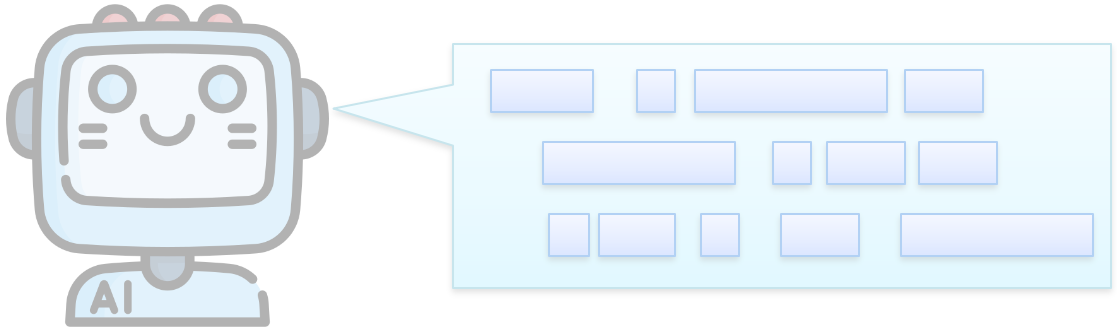
Extract the model information by querying the model in a black-box setting



I can obtain a similar model to yours at a much lower cost

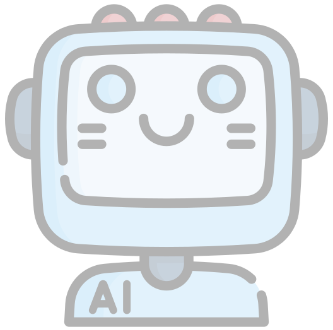
Can we watermark the model?

Text Watermark



Is this text generated by my model?

Model Watermark



Part III: Model Watermarking

s model from
y model?

Outline

- Part I: Introduction
- Part II: Text Watermark
  - (a) Green-Red Watermark
  - (b) Gumbel Watermark
  - (c) Theoretical results
- Part III: Model Watermark
- Part IV: Conclusion and Future Directions

NEURAL INFORMATION PROCESSING SYSTEMS

Website; Q&A

Watermarking for Large Language Models

Part II: Text Watermarking

Xuandong Zhao
UC Berkeley

Yu-Xiang Wang
UC San Diego

Lei Li
CMU

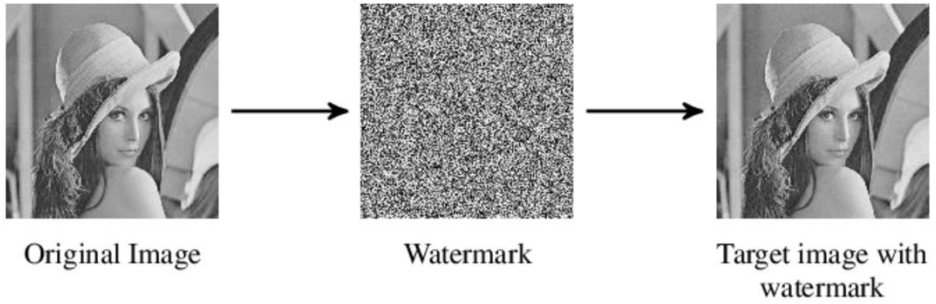
Watermarking has a long history

The *Crown CA* watermark found on many British Commonwealth stamps

https://en.wikipedia.org/wiki/Watermark

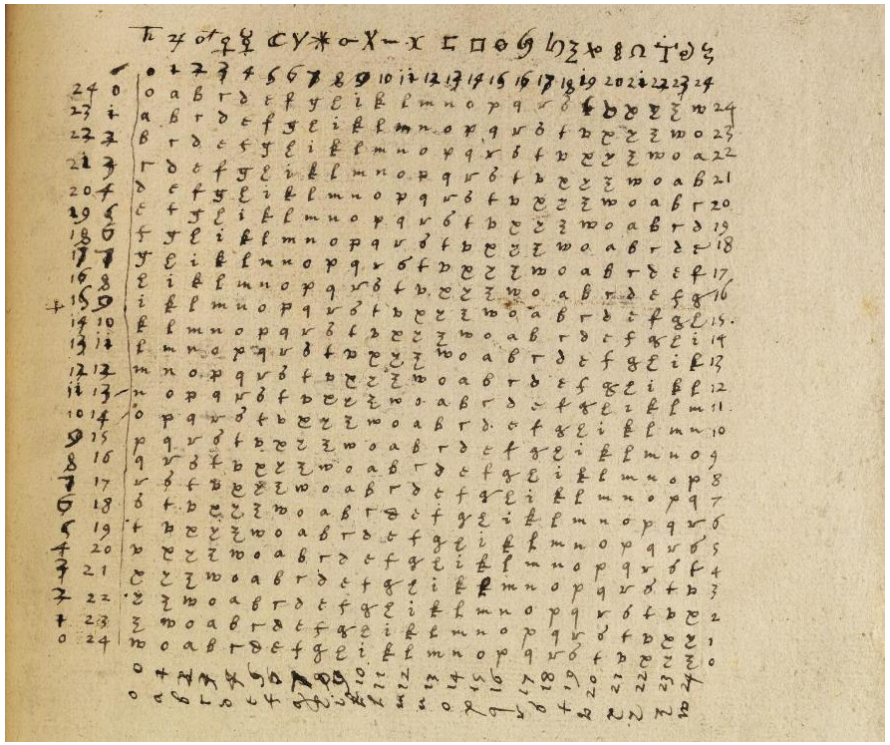
Traditional Image Watermarks

Invisible Image Watermarks



Text Watermarking

- Ancient Greece: Steganography
- 1499: Trithemius “Steganographia”



https://en.wikipedia.org/wiki/Steganography

- 1950s: Embedding code to music (Hembrooke, 1954)
- 1990s to 2000s: Digital Watermarks (e.g., Ingemar J. Cox, Matt Miller, etc..)
- Rule-based parsed syntactic tree (Atallah et al., 2001)
- Rule-based semantic structure of text (Atallah et al., 2000; Topkara et al., 2006)
- Neural steganography with DL models (Fang et al., 2017; Ziegler et al., 2019)

2022+: Recent Renaissance due to the rise of Generative AI

- Watermarking LLM text

  Aaronson (2022), Kirchenbauer et al. (2023), **Zhao et al. (2023; 2024),** Christ et al. (2023), Kuditipudi et al. (2023), Hu et al. (2023), Christ and Gun (2024)

  Part 2 of the tutorial

- Watermarking LLM models

  Zhao et al. (2022) "Distillation resistant watermarking", **Zhao et al. (2023)** "Protecting Language Generation Models via Invisible Watermarking"

- Watermarking Images (e.g. from Diffusion models)

  E.g., Fernandez et al. (2023) "Stable signature", Wen et al. (2023) "Tree-Ring Watermarks"

- "Is strong watermarking possible?" – Watermark attacks

  Zhao et al. (2023) "Invisible Image Watermarks Are Provably Removable Using Generative AI"

  Zhang, Barak et al. (2024) "Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models"

  Sadasivan et al. (2023) "Can AI-generated text be reliably detected?"

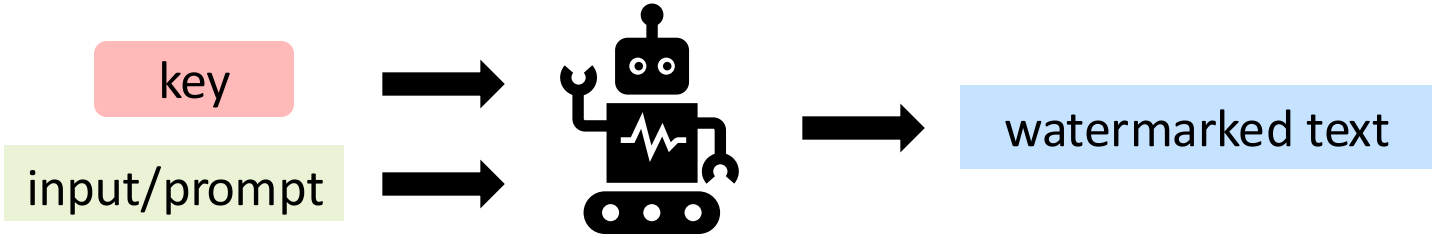
Slightly different settings, motivating applications and new challenges.

Main Difference

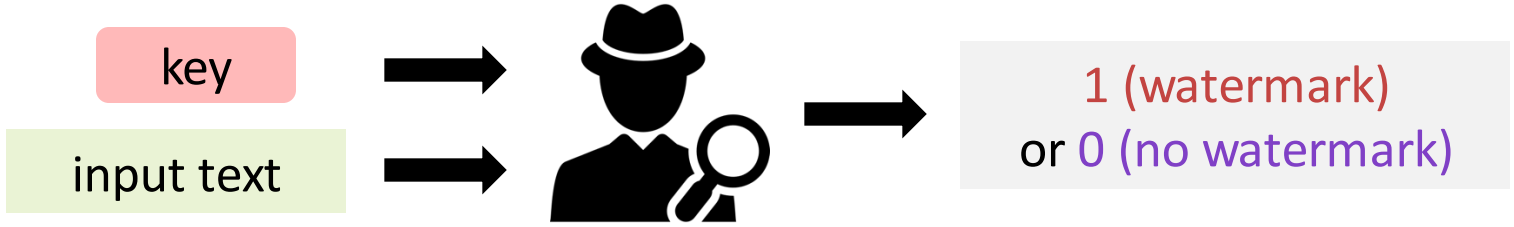
- Steganography / Watermarking in the 1990s to 2000s
  - We are given the text / image to be protected.
- Modern GenAI watermarks
  - We also have access to the generative process.

An LM Watermarking Scheme has two components

- **Watermark**($\mathcal{M}$): (possibly randomized procedure) that outputs a new model $\hat{\mathcal{M}}$, and detection key k



- **Detect**($k, \boldsymbol{y}$): takes input detection key k and sequence $\boldsymbol{y}$, then outputs 1 (indicating it was AI-generated) or 0 (indicating it was human-generated)



Desired Properties of an Ideal Watermark

- **Quality of Generated Text**

- **Detection Accuracy Guarantee**

  - Type I error: "No false positives" → won't catch human text
  - Type II error: ext

We will have a detailed discussion later.
@Yu-Xiang Wang

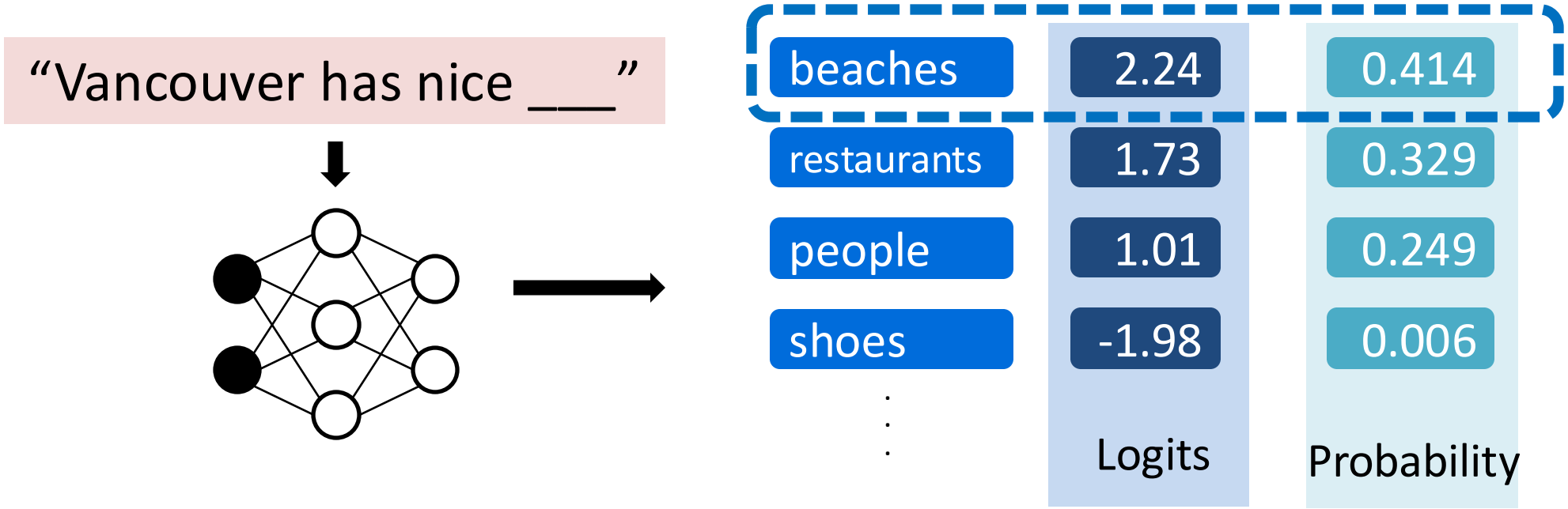
Check paper "SoK: Watermarking for AI-Generated Content" for more details

- **Robust**
  - Be robust against evasion attacks, e.g., post-editing.
- **Unforgeability (Security) Guarantee**

  - Can not easily produce wm content w/o the wm key.

What is a Language Model?

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

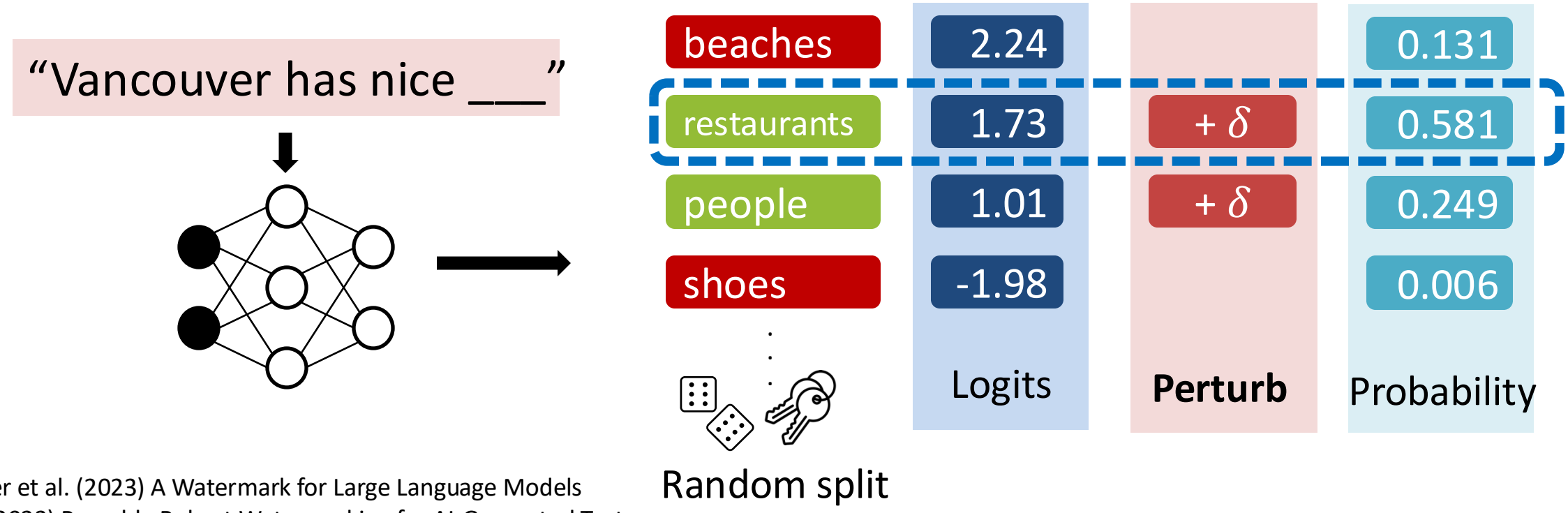


The **universe of words** is called a **vocabulary** V

Green-Red Watermark

(Kirchenbauer et al. 2023; Zhao et al. 2023)

$\hat{\mathcal{M}}$: Modified LM
Key: Green lists
Detection: Count # of Greens



Kirchenbauer et al. (2023) A Watermark for Large Language Models
Zhao et al. (2023) Provable Robust Watermarking for AI-Generated Text

Green-Red Watermark

(Kirchenbauer et al. 2023; Zhao et al. 2023)

$\mathcal{M}$: $y_t \sim$ Softmax(logits(Prompt, $y_{<t}$))

$\widehat{\mathcal{M}}$: $y_t \sim$ Softmax(logits(Prompt, $y_{<t}$) + $\delta \cdot \mathbf{1}(\cdot\ is\ green)$)

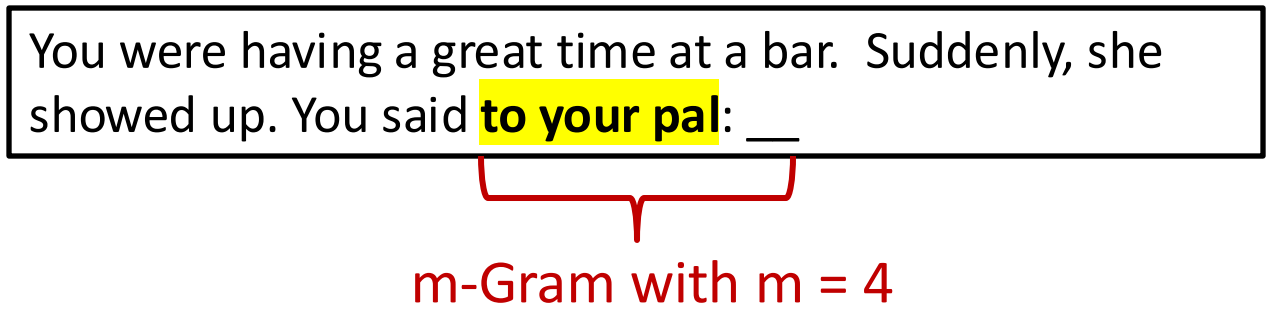
Increase the probability of green tokens slightly.

Decrease the probability of red tokens slightly.

Kirchenbauer et al. (2023) A Watermark for Large Language Models
Zhao et al. (2023) Provable Robust Watermarking for AI-Generated Text

How is the *Green* list generated?

- ***Randomly*** **selecting** γ fraction of the vocabulary, e.g., 0.5
- (Kirchenbauer et al. [KGW-Watermark]): Different green list at each time t as function of the prefix with length (m-1). Default: m=2



- (Zhao et al. [Unigram-Watermark]): Use m = 1, i.e., a consistent "Green list".

Detection of Green-Red Watermark

Input: Suspect text $y = [y_1, \ldots, y_n]$, e.g. "Over the …"

(Optional pre-processing) $y = \text{unique}(y)$

1. Compute the ***z*-score**:

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

Num of Green tokens

2. If $z >$ *threshold* then
   Return "y is watermarked"
   Else
   Return "no evidence"

Green-Red Watermark Examples

LLaMA-13B, unwatermarked
z-score=-2.4

Prompt: Can I succeed after many failures?

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...]

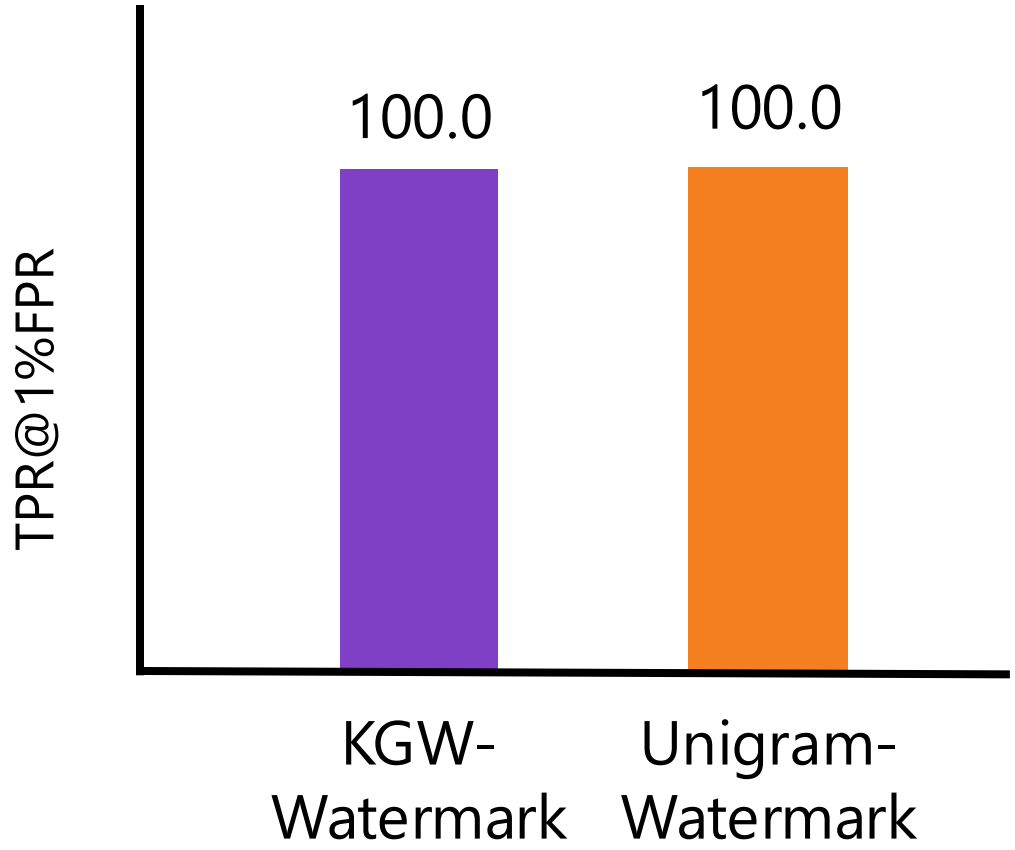
LLaMA-13B, watermarked
z-score=11

Prompt: Can I succeed after many failures?

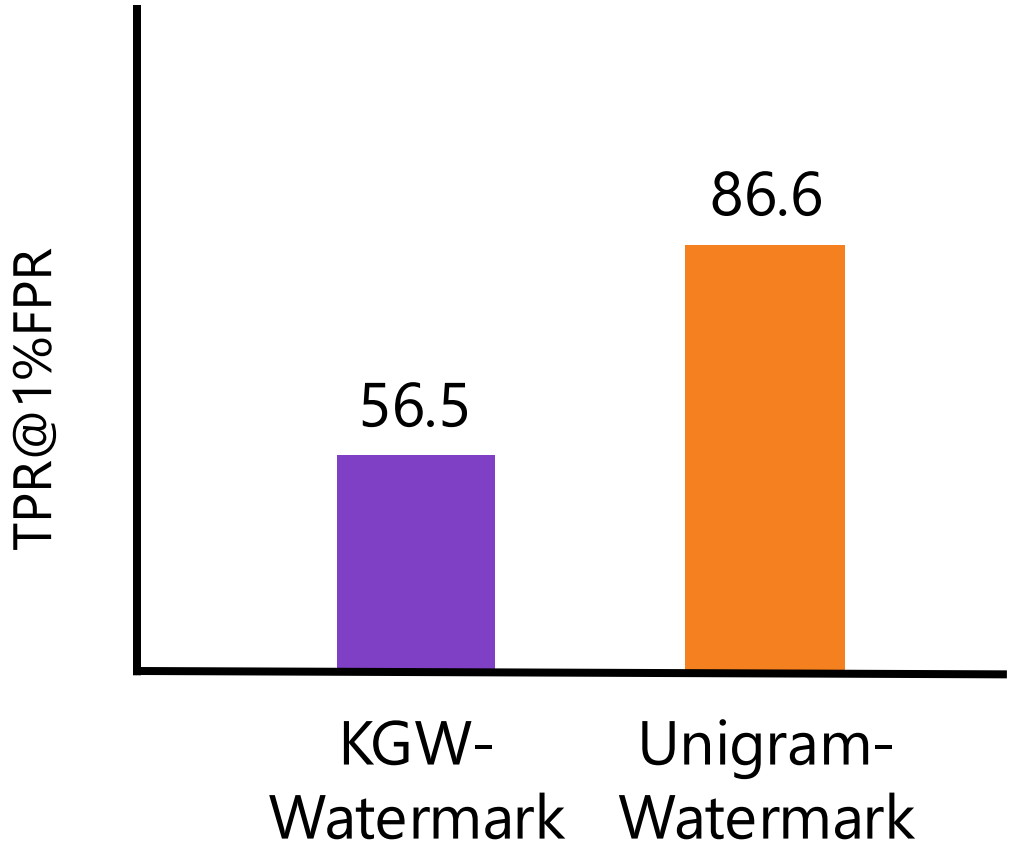
A: When most people are confronted with failure, they cannot imagine such a thing happening. When one faces business reverses and bankruptcy, it seems impossible. When we are rejected it looks as if we are going to be rejected forever. However, it does not need to be this way. The human spirit simply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

Let us try a live demo!

Empirical Results

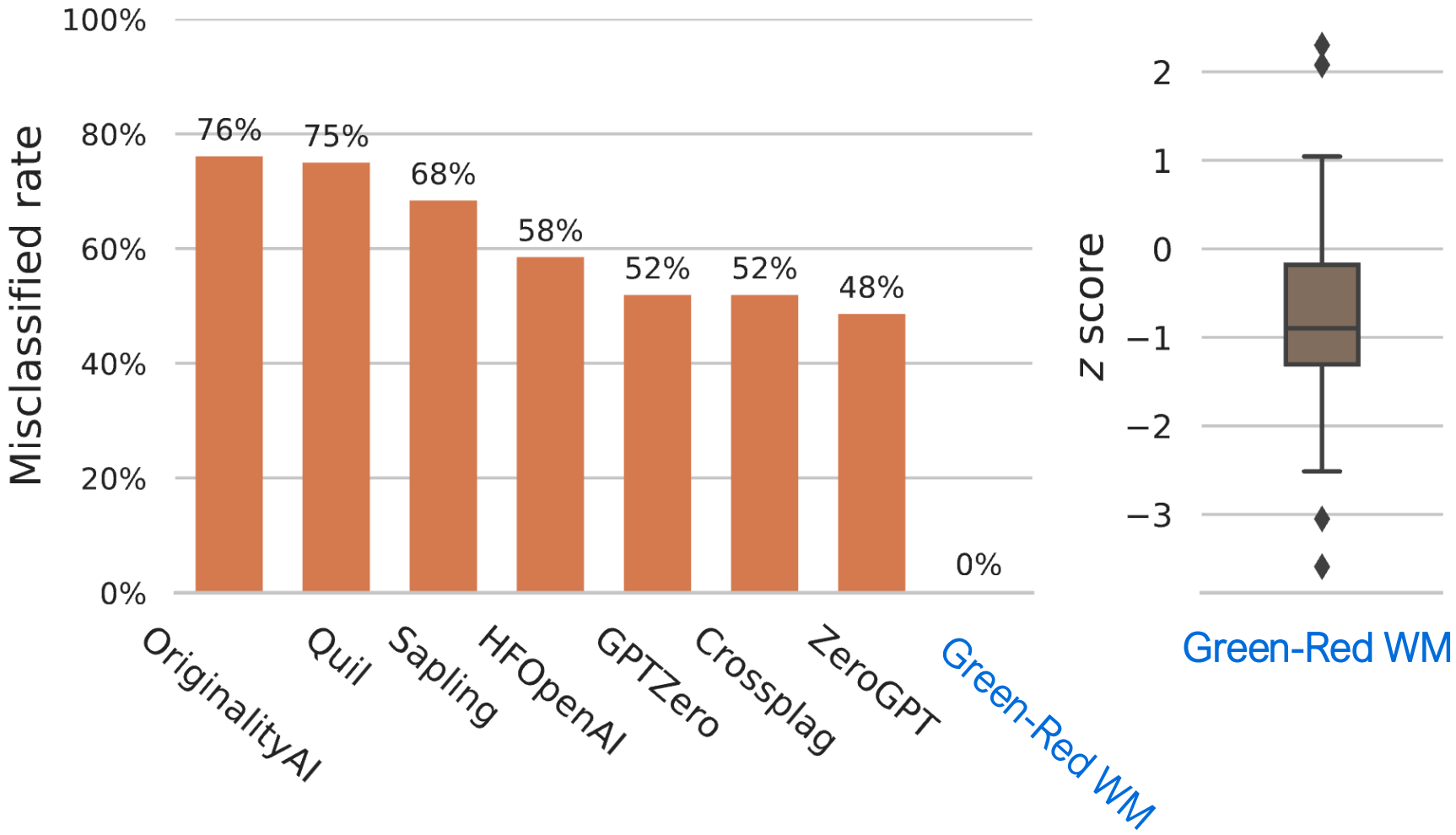


No Attack



ChatGPT Paraphrasing

Empirical Results



Distinguishing human-written text on TOEFL dataset (Out of distribution)

Different versions of Green-Red WM

- Green-Red watermark for code generation (Lee et al., 2023; Guan et al., 2024)
- Adaptive/dynamic perturbations in the logits (Liu et al., 2023; Huo et al., 2024, Liu et al., 2024)
- Public key (Liu et al., 2023; Zhou et al., 2024)
- Multi-bits (Yoo et al., 2023; Fernandez et al., 2023)
- Many others…

Yu-Xiang will provide more in-depth details!