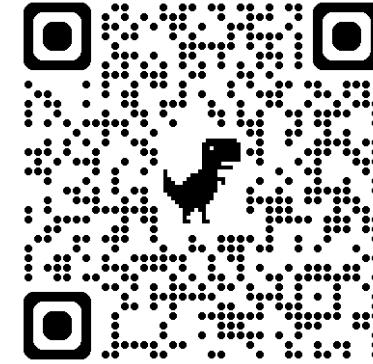


ACL 2024

Bangkok, Thailand



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

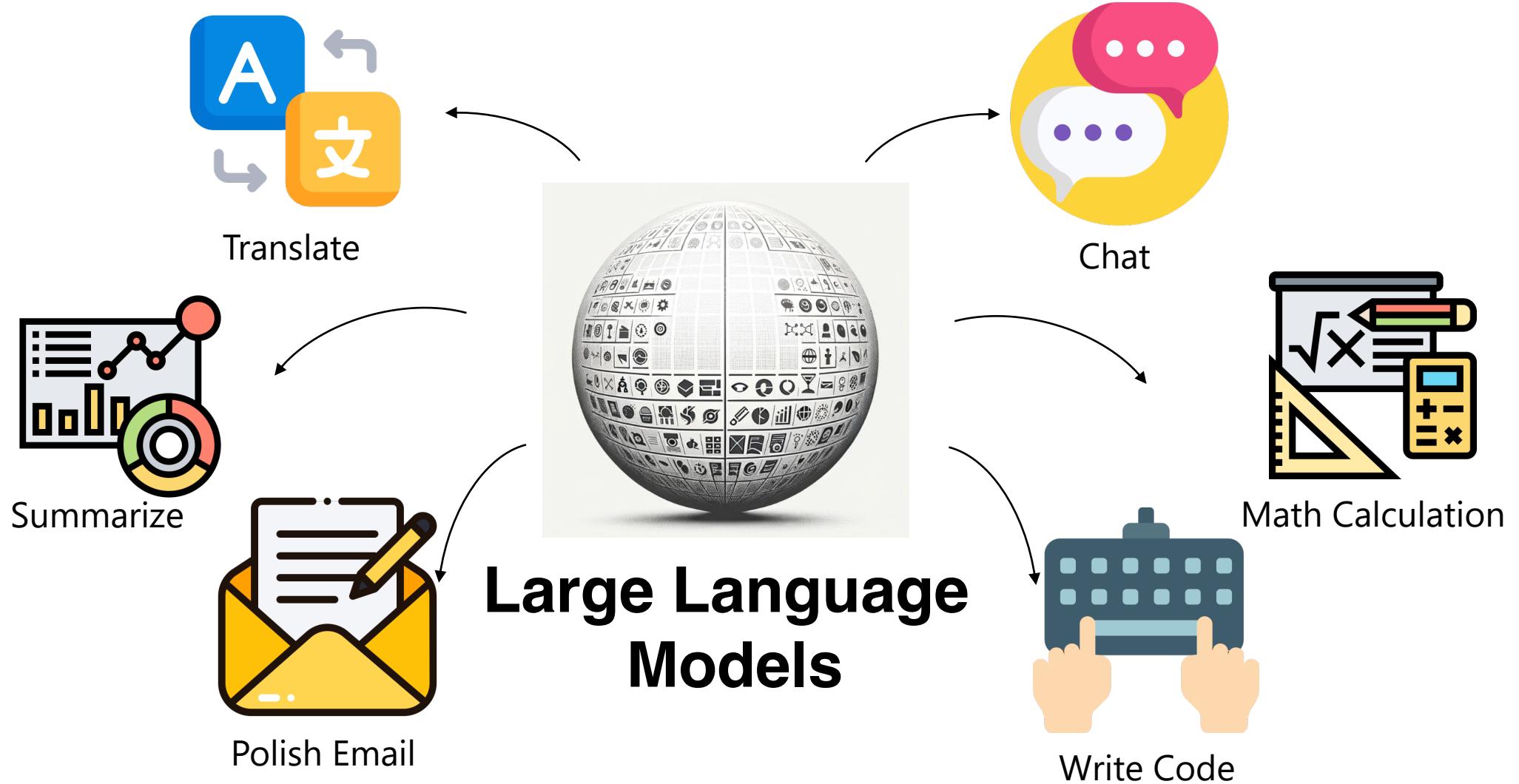
2023

2024



Supercharge your creativity
and productivity

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



How people use LLMs/Chatbots?

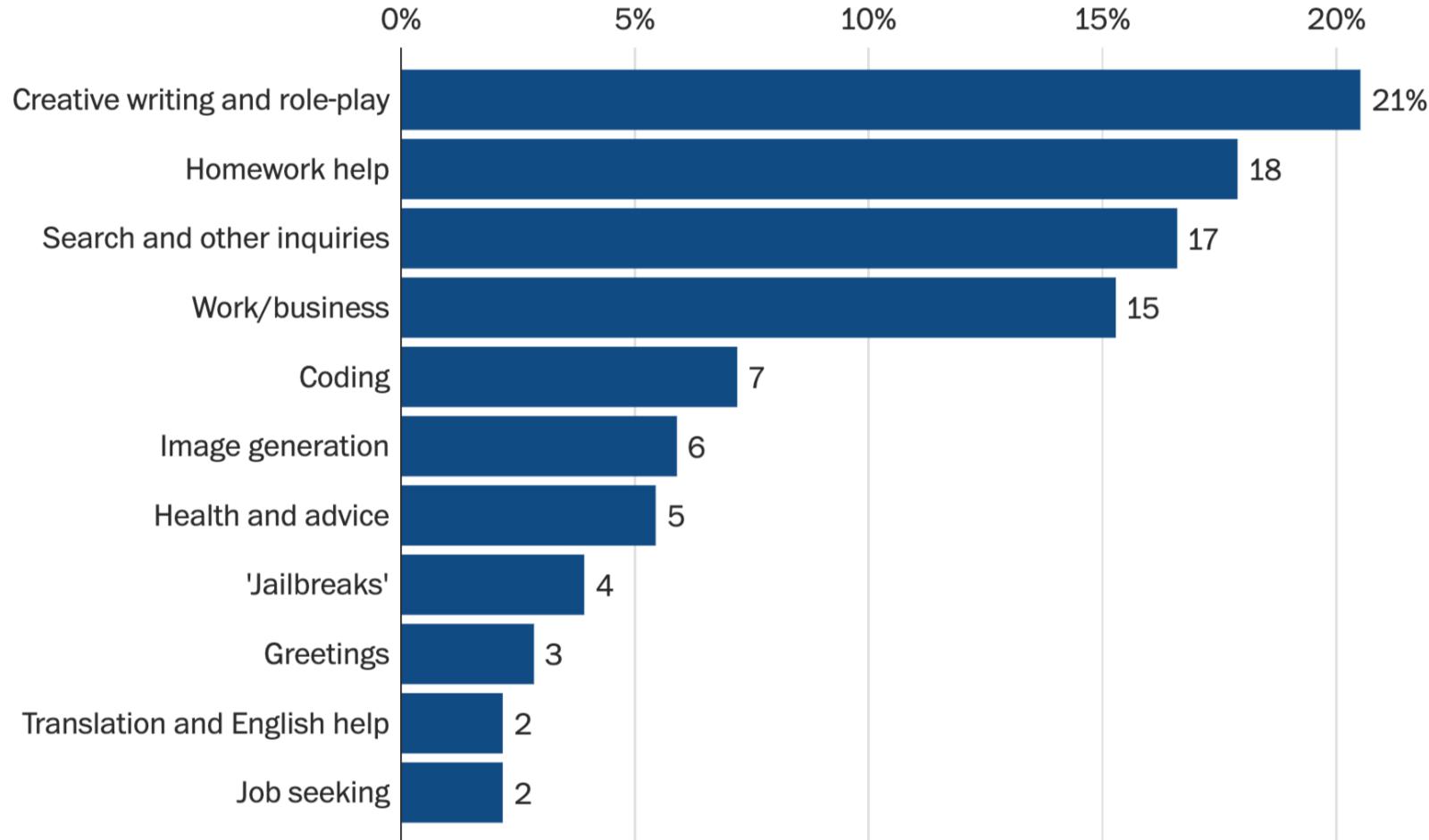


Chart shows proportion of prompts in the category from a random sample of 458 English WildChat conversations, selected from the first prompt per day per US-based IP address. Margin of sampling error is 5 percentage points.

Risks of LLMs

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



China reports first arrest over fake news

BREAKING

Judge Fines Two Lawyers For Using

Artificial Intelligence

ChatGPT Leaks Se... User Data, OpenAI Suspects

The leaks exposed conversations, per

Chris Westfall Contributor ⓘ
Guidance for leaders and aspiring leaders, interested in career impact

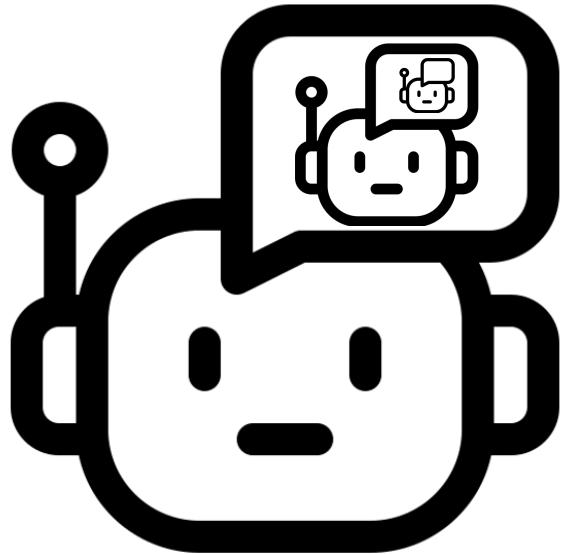
Forbes

Follow

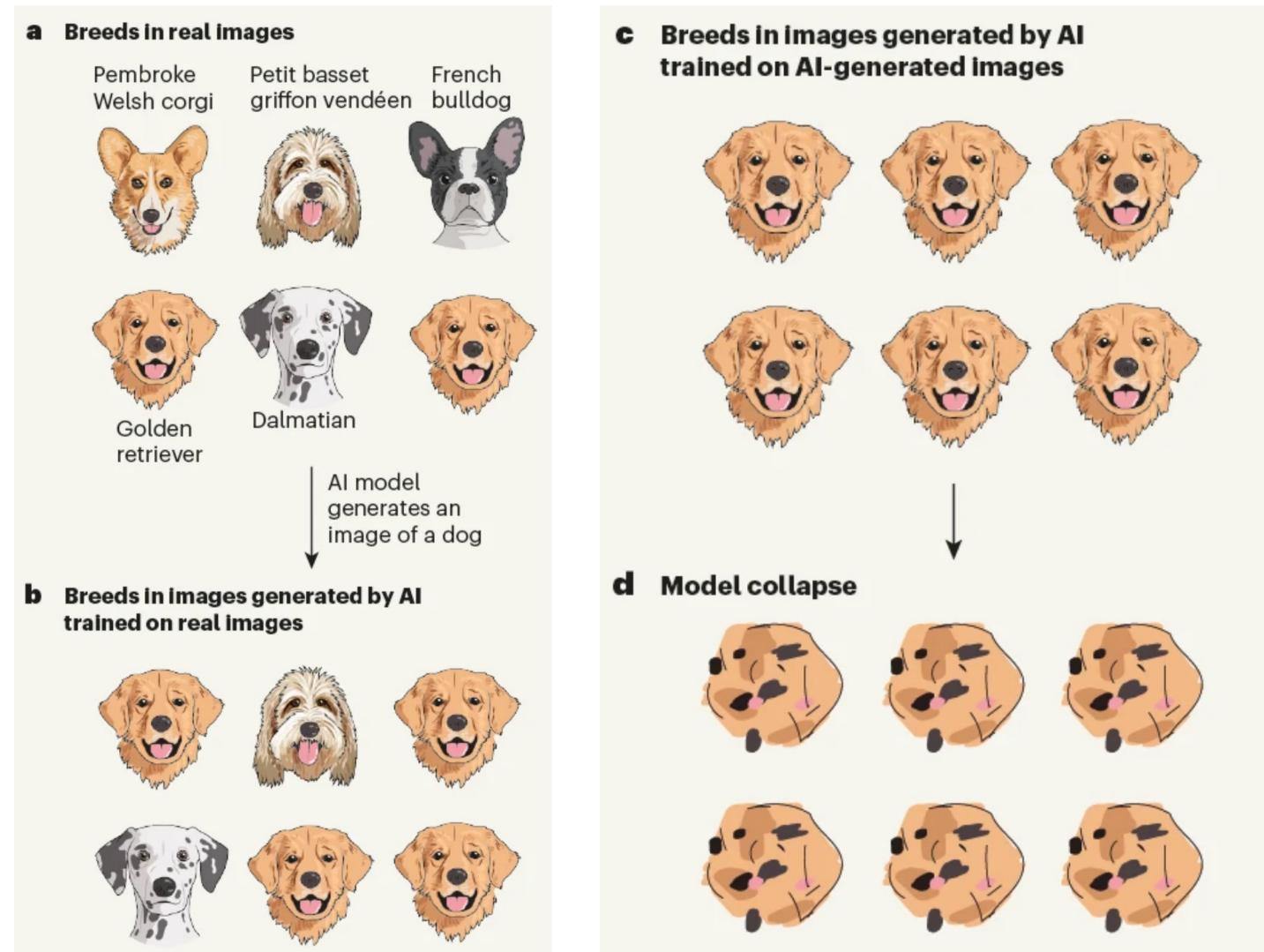
Why do we need to detect AI-generated text?



Why do we need to detect AI-generated text?



**Model Degeneration
or
Model Collapse**



Why do we need to detect AI-generated text?



≡ Google DeepMind

Executive
and Trusty
Use of

means us
managin



dible speakers who will be sha
leaders in their field and hav
s, we will also have other enga
t sessions and networking op
portunity to dive deeper into
nships.

TECHNOLOGY

great success, and I'd love to

SynthID

Identifying AI-generated content with SynthID

Share



That means
n or has
ote bias and
Finally, it
ange, and



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.

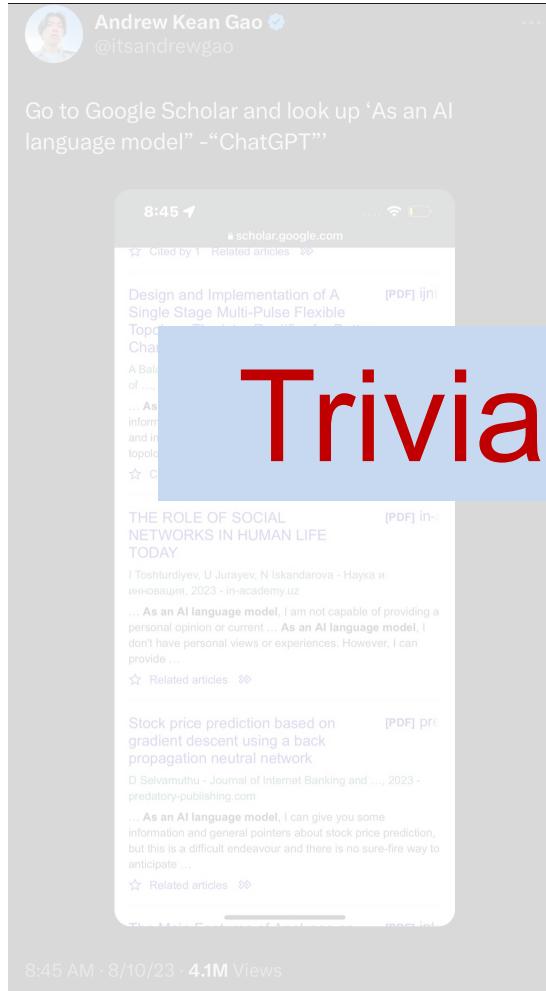
Over the river, and through the wood,
Oh, how the wind does blow!
It stings the toes
And bites the nose
As over the ground we go.



Child, Lydia Maria. "Thanksgiving Day." 1844.

How to detect AI-generated text?

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

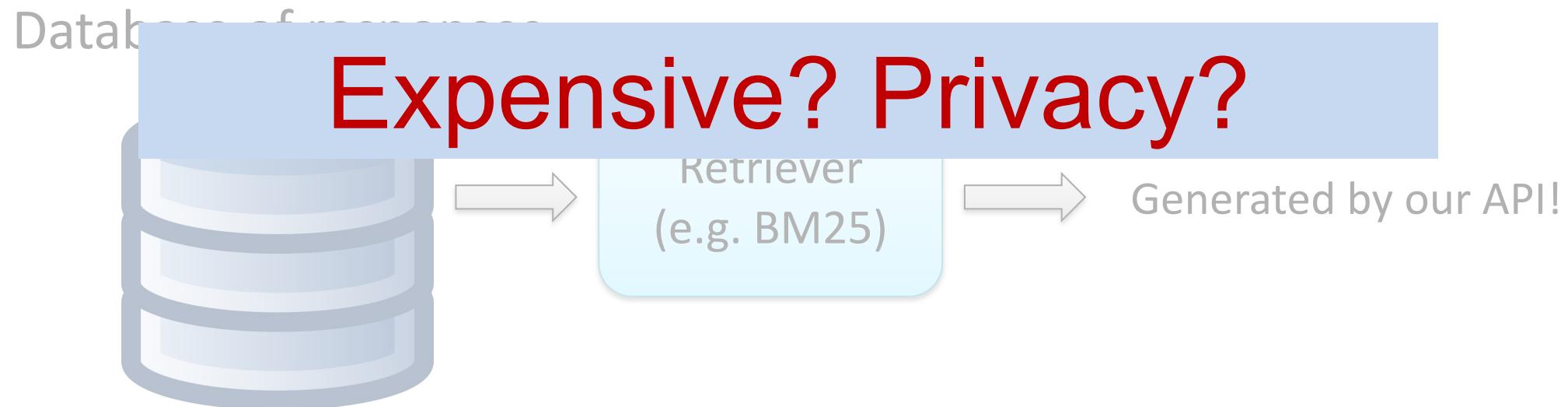


III. PROPOSED SYSTEM

As an AI language model, I can provide some general information on the proposed system for the analysis, design, and implementation of a single-stage multi-pulse flexible-topology thyristor rectifier for battery charging in electric vehicles. The proposed system aims to develop a high-efficiency and reliable battery charging system for electric vehicles. The system utilizes a single-stage multi-pulse flexible-topology thyristor rectifier to ... This rectifier has a flexible ...
...
mechanism regulates the rectifier's output voltage and current to ensure that the battery is charged in a controlled and safe manner. It also monitors the battery's status and adjusts the charging process accordingly to prevent overcharging or overheating. The proposed system is designed to be compact and lightweight, making it easy to install and use in electric vehicles. It is also designed to be cost-effective while providing high performance and reliability, which is essential for the widespread adoption of electric vehicles. The proposed system for the analysis, design, and implementation of a single-stage multi-pulse flexible-topology thyristor rectifier for battery charging in electric vehicles is a promising solution for the development of efficient and reliable battery charging systems for electric vehicles.

How to detect AI-generated text?

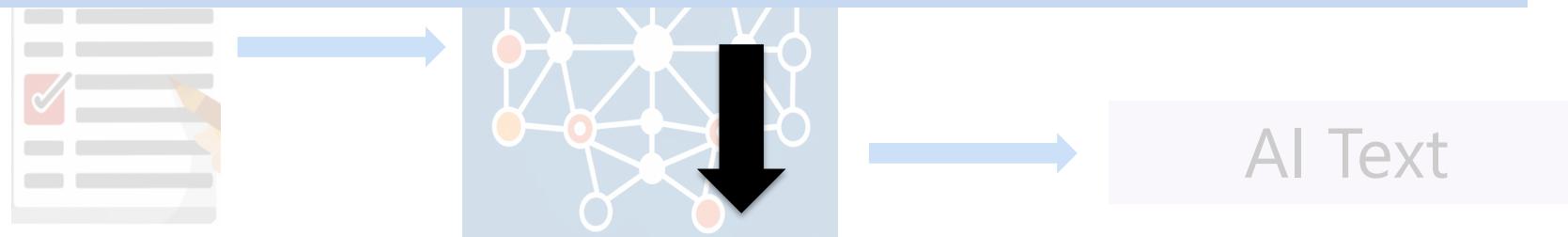
- Maintain a database of all completions



How to detect AI-generated text?

- Train classification models [GPTZero, Turnitin, ...]

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



Part IV: Post-Hoc Detection

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

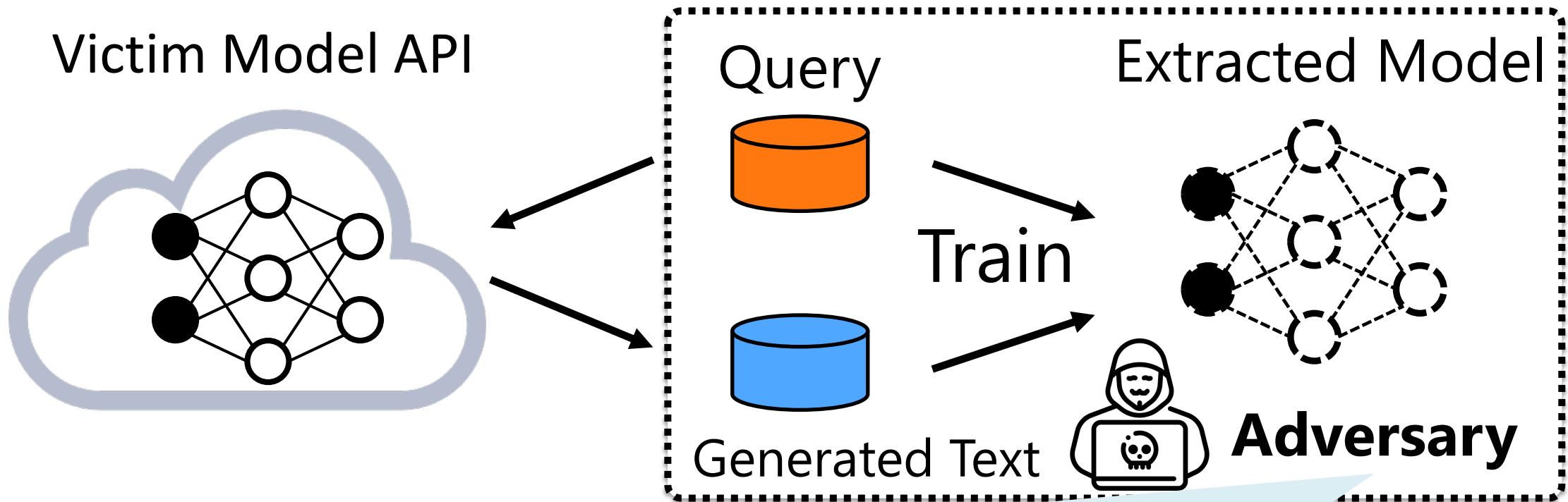
Intellectual Property of LLM



I want to steal the
model!

Model Stealing/Extraction Attack

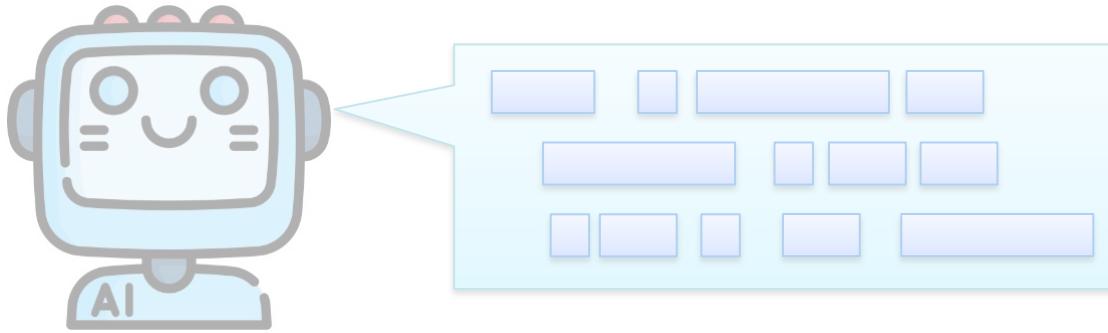
Extract the model information by querying the model in a black-box setting



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

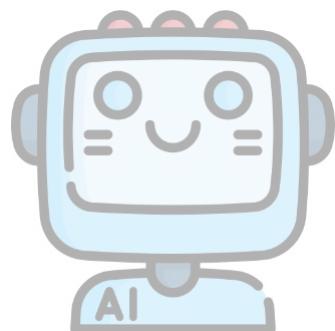
Can we watermark the model?

Text Watermark



Is this text generated
by my model?

Model Watermark



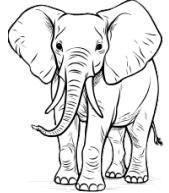
Is this model from
my model?

Outline

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

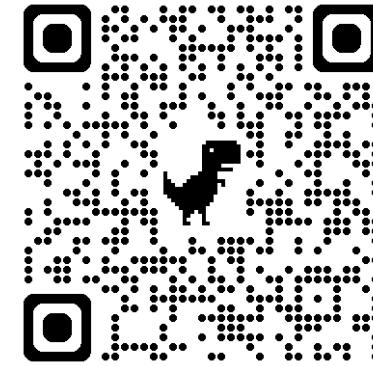
Break





ACL 2024

Bangkok, Thailand



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



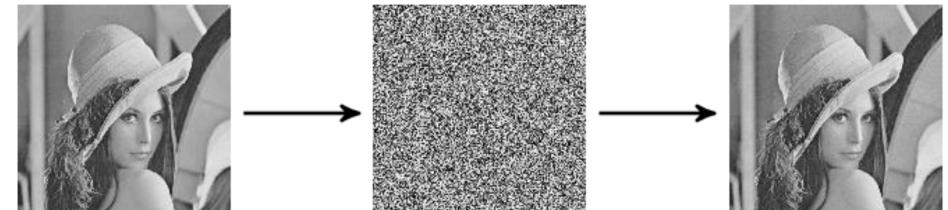
Traditional Image Watermarks

Invisible Image Watermarks



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

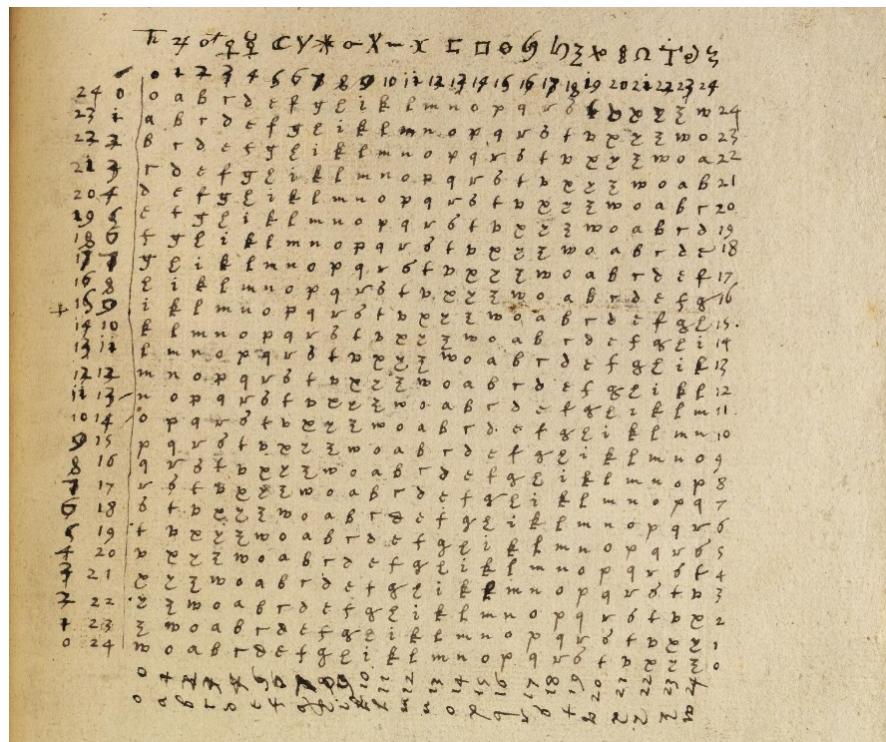
<https://en.wikipedia.org/wiki/Watermark>



Target image with
watermark

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?”

“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?

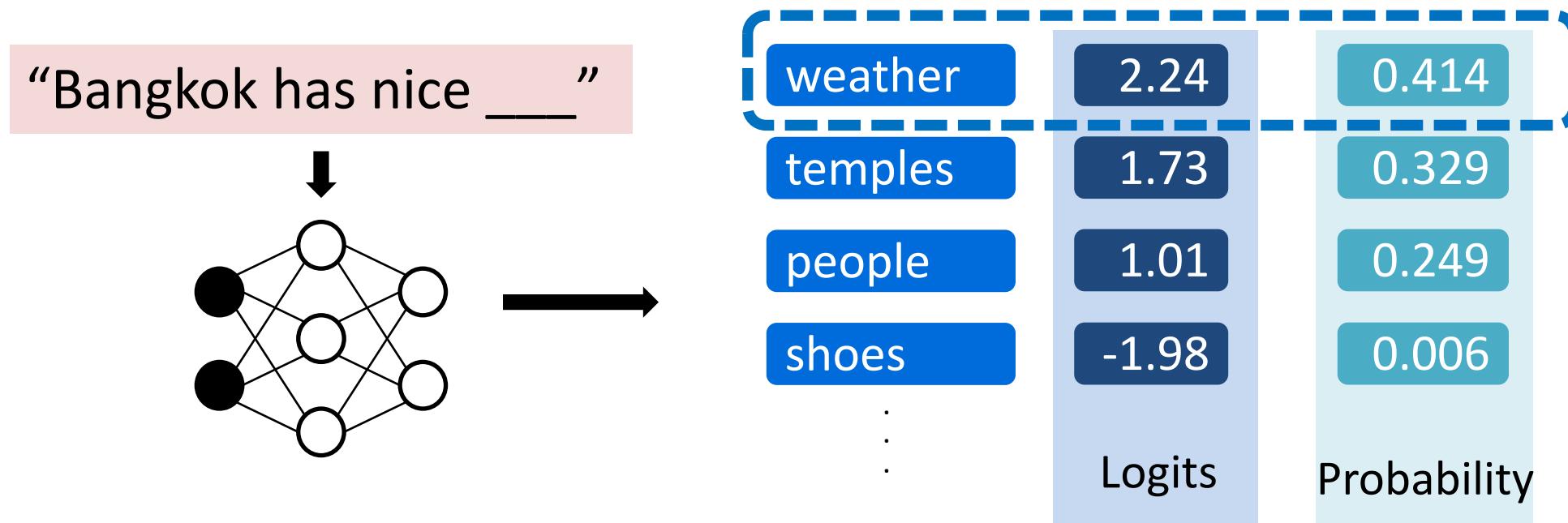
Slightly different settings, motivating applications and new challenges.

Main difference

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

What is a Language Model?

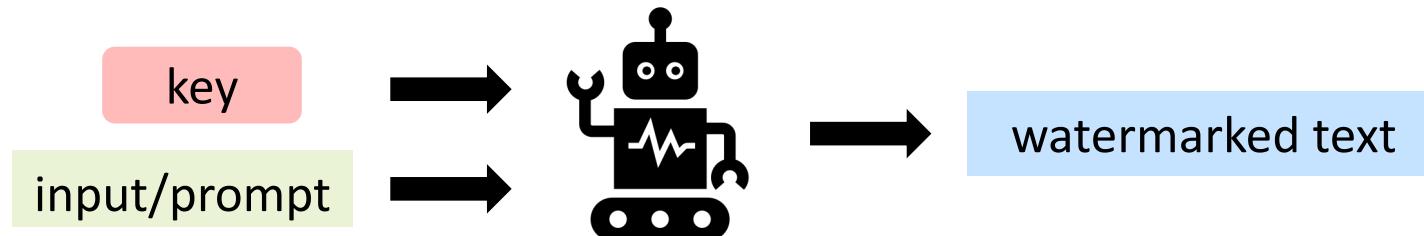
$P(\text{next word } y_t \mid \text{Prompt } x, \text{ previous words } y_{1:t-1})$



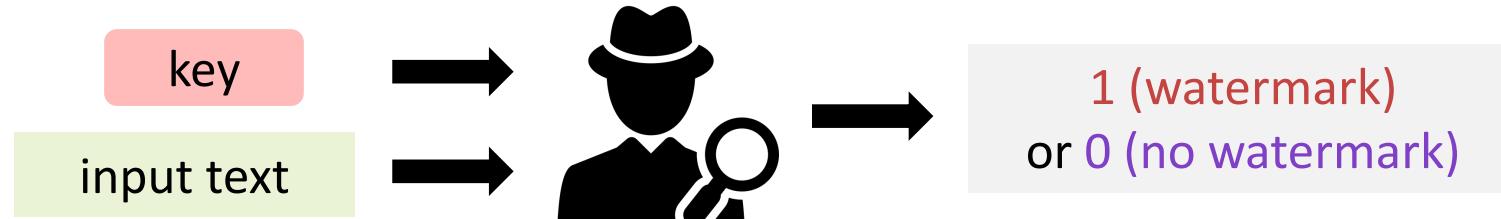
The **universe of words** is called a **vocabulary V**

An LM Watermarking Scheme has two components

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



- $\text{Detect}(k, y)$: takes input detection key k and sequence 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: “No false negatives” → miss real watermark text

- **Robustness**

@Yu-Xiang Wang

- Be robust against evasion attacks, e.g., post-editing.

- **Security Guarantee**



- Can not easily guess the watermark key.

Green-Red Watermark

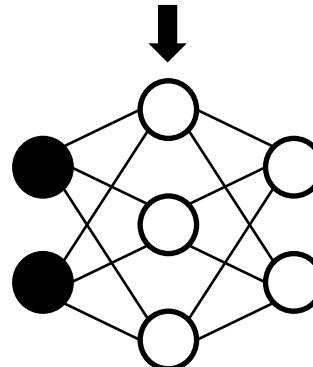
(Kirchenbauer et al. 2023; Zhao et al. 2023)

$\hat{\mathcal{M}}$: Modified LM

Key: Green lists

Detection: Count # of Greens

““Bangkok has nice ___”



weather

temples

people

shoes



2.24

1.73

1.01

-1.98

Logits

+ δ

+ δ

0.131

0.581

0.249

0.006

Perturb

Probability

Random split

Green-Red Watermark

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

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

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

Increase the probability of green tokens slightly.

Decrease the probability of red tokens slightly.

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: $m=2$

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



m-Gram with $m = 4$

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

Detection of Green-Red Watermark

Input: Suspect text $y = [y_1, \dots, y_n]$, e.g. “Over the ...”

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

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

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

2. If $z > \text{threshold}$ then

 Return “y is watermarked”

Else

 Return “no evidence”

Num of **Green tokens**

Green-Red Watermark Examples

Prompt: Can I succeed after many failures?

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

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...]

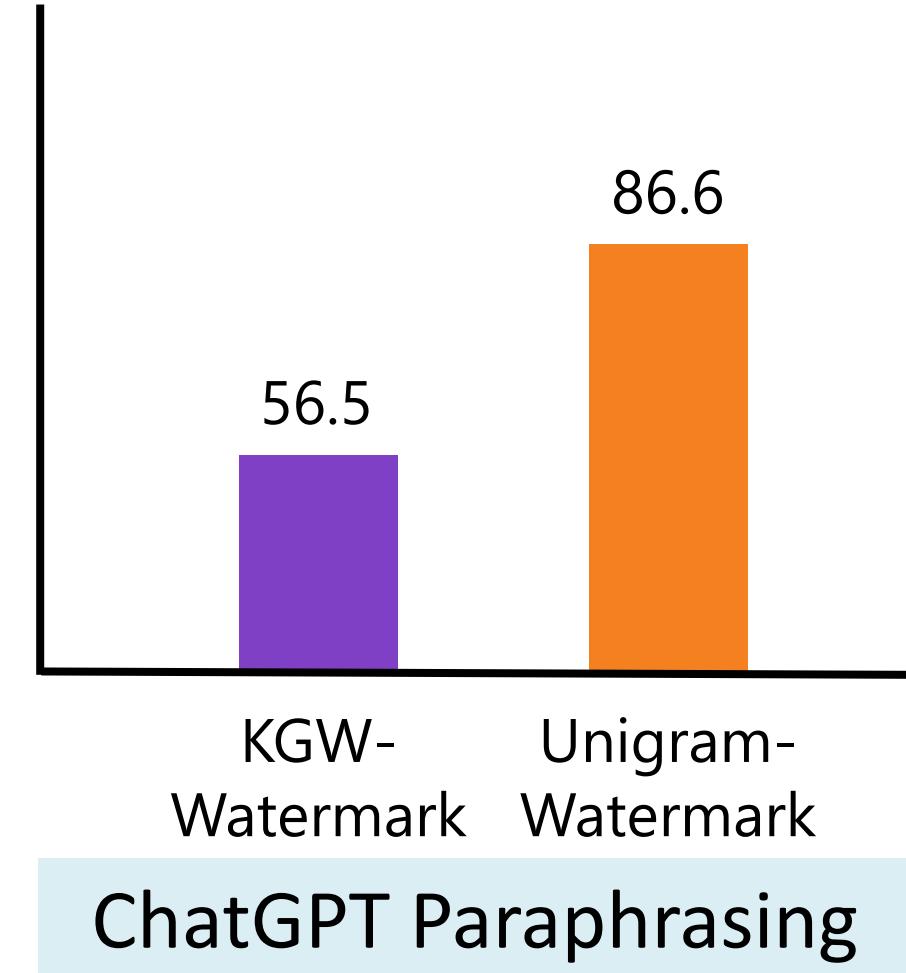
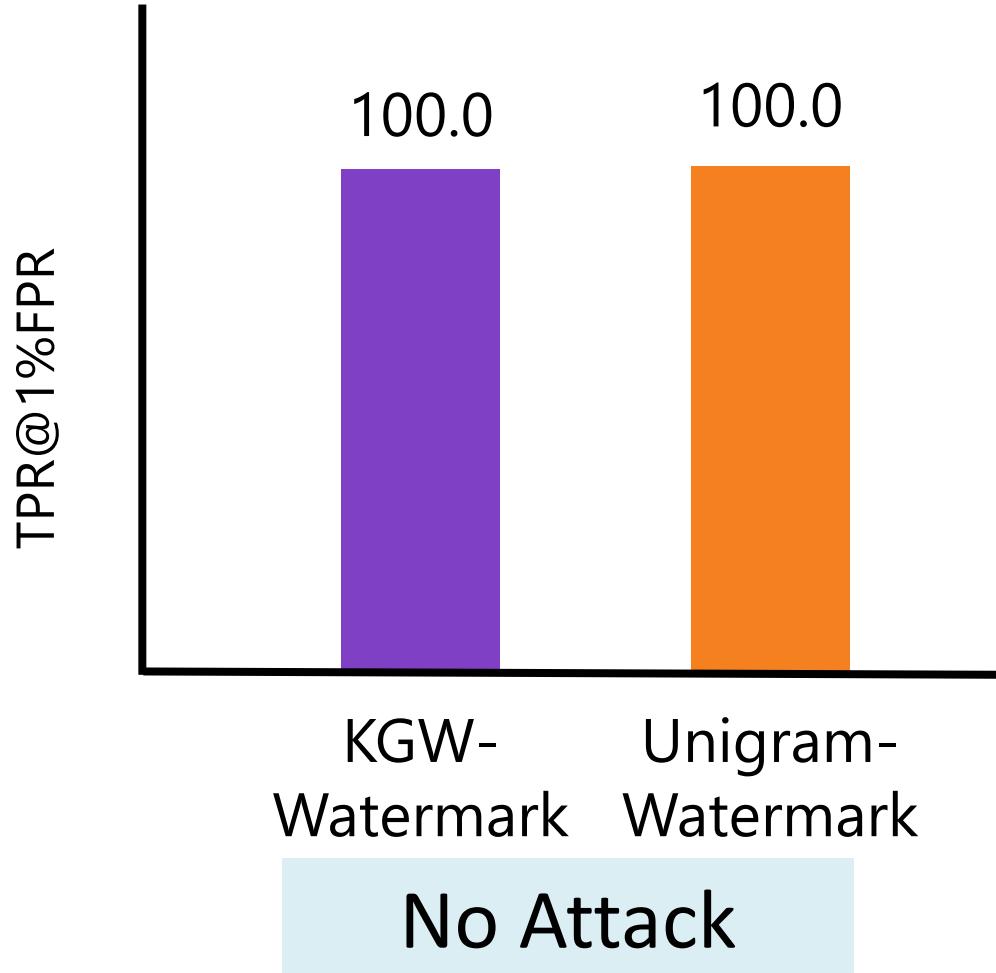
Prompt: Can I succeed after many failures?

LLaMA-13B, watermarked
z-score=11

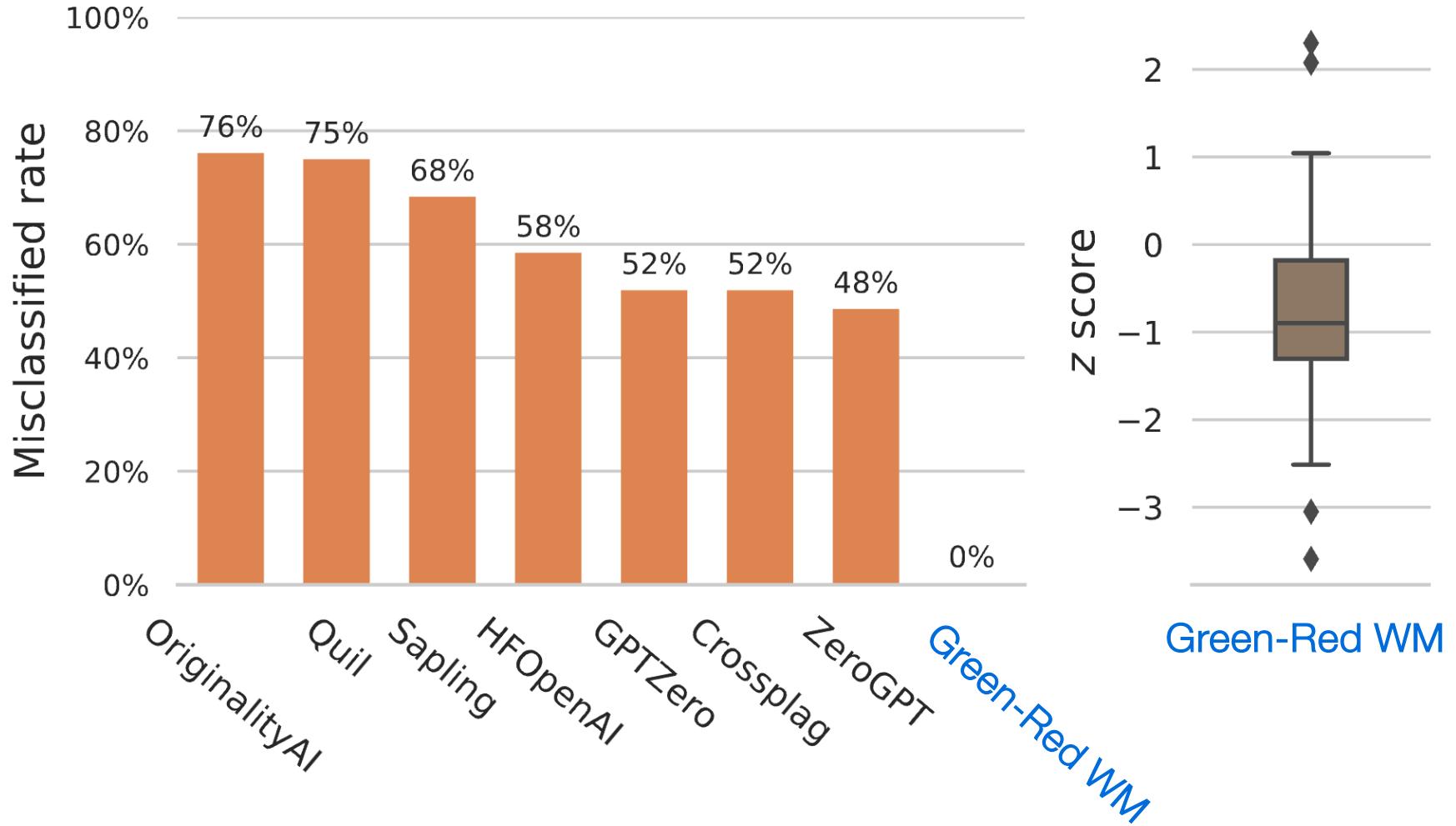
A: When most people are confronted with failure, they cannot imagine such a thing happening. When one faces business reverses and bankruptcy, it seems impossible. When we are rejected it looks as if we are going to be rejected forever. However, it does not need to be this way. The human spirit simply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

Let us try a live demo!

Empirical Results



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!