

Watermarking and Detecting AI Generation

Lei Li
Language Technologies Institute
School of Computer Science
Carnegie Mellon University

November 3, 2024

NN LM

BERT

2000

2018



2020

2023

2024



Google Gemini

Supercharge your creativity
and productivity

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

What is different before/after LLM?

- LLM is an Omniscient
 - give decent answers to almost all common topics
- LLM drives productivity
 - translate, edit, summarize, calculate, write code
- LLM as idea generator
 - suggest paper title, recommend a restaurant, create a poem

How people use LLMs/Chatbots?

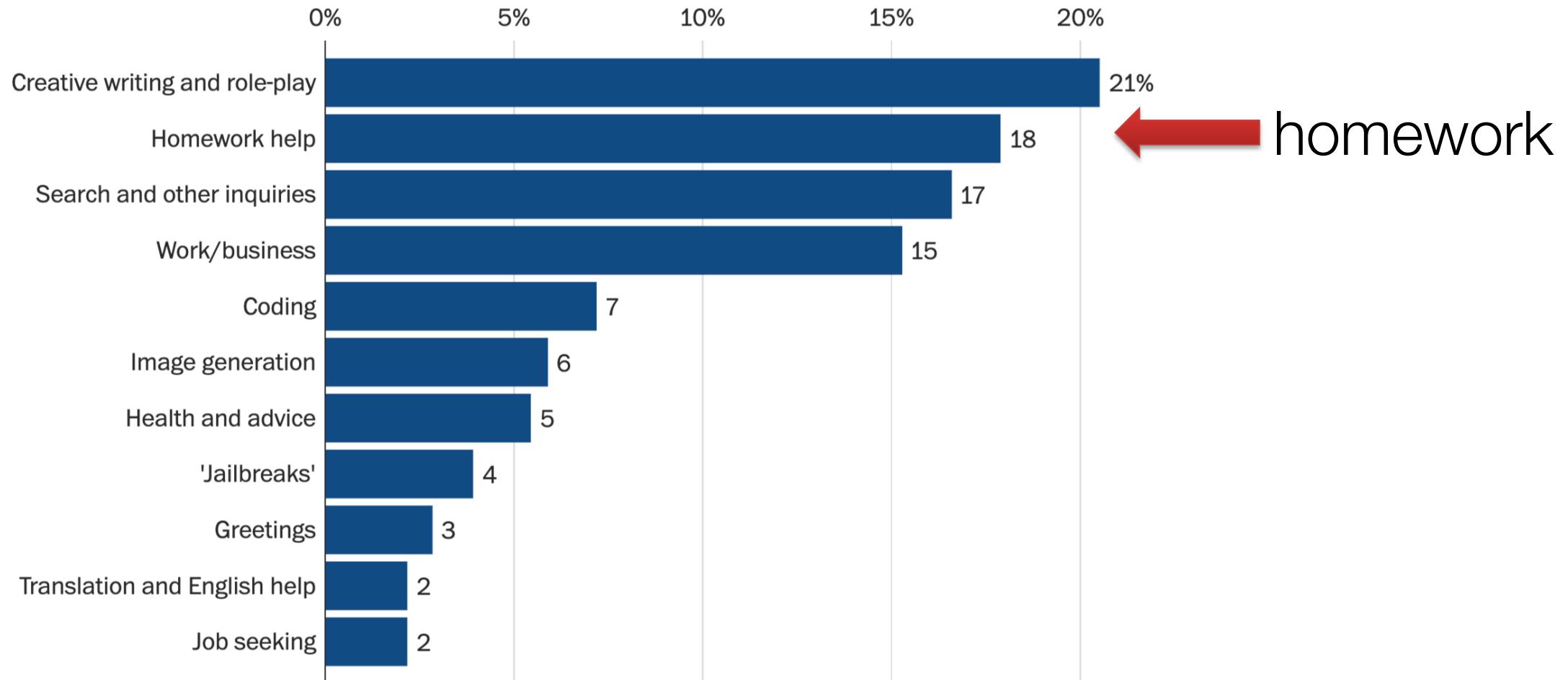


Chart shows proportion of prompts in the category from a random sample of 458 English WildChat conversations, selected from the first prompt per day per US-based IP address. Margin of sampling error is 5 percentage points.

Risks of LLMs

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

The image is a composite of two visual elements. On the left, a vertical list of bullet-pointed risks is displayed against a white background. On the right, a news article snippet from 'FORTUNE' magazine is shown, featuring a large yellow warning triangle with an exclamation mark over a headline about AI leaks.

≡ SEARCH

FORTUNE

SIGN IN

Subscribe Now

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

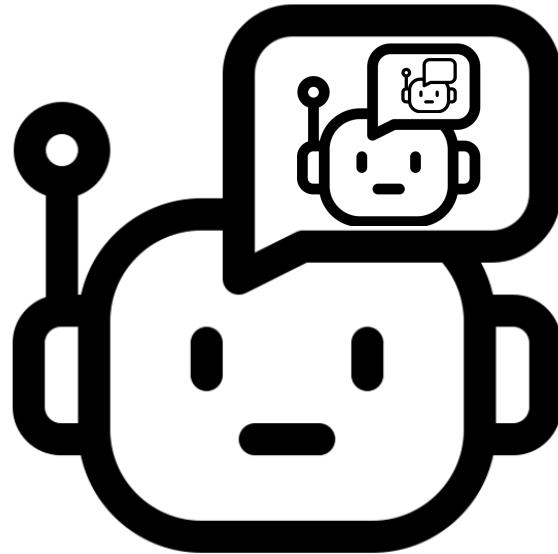
Forbes

Follow

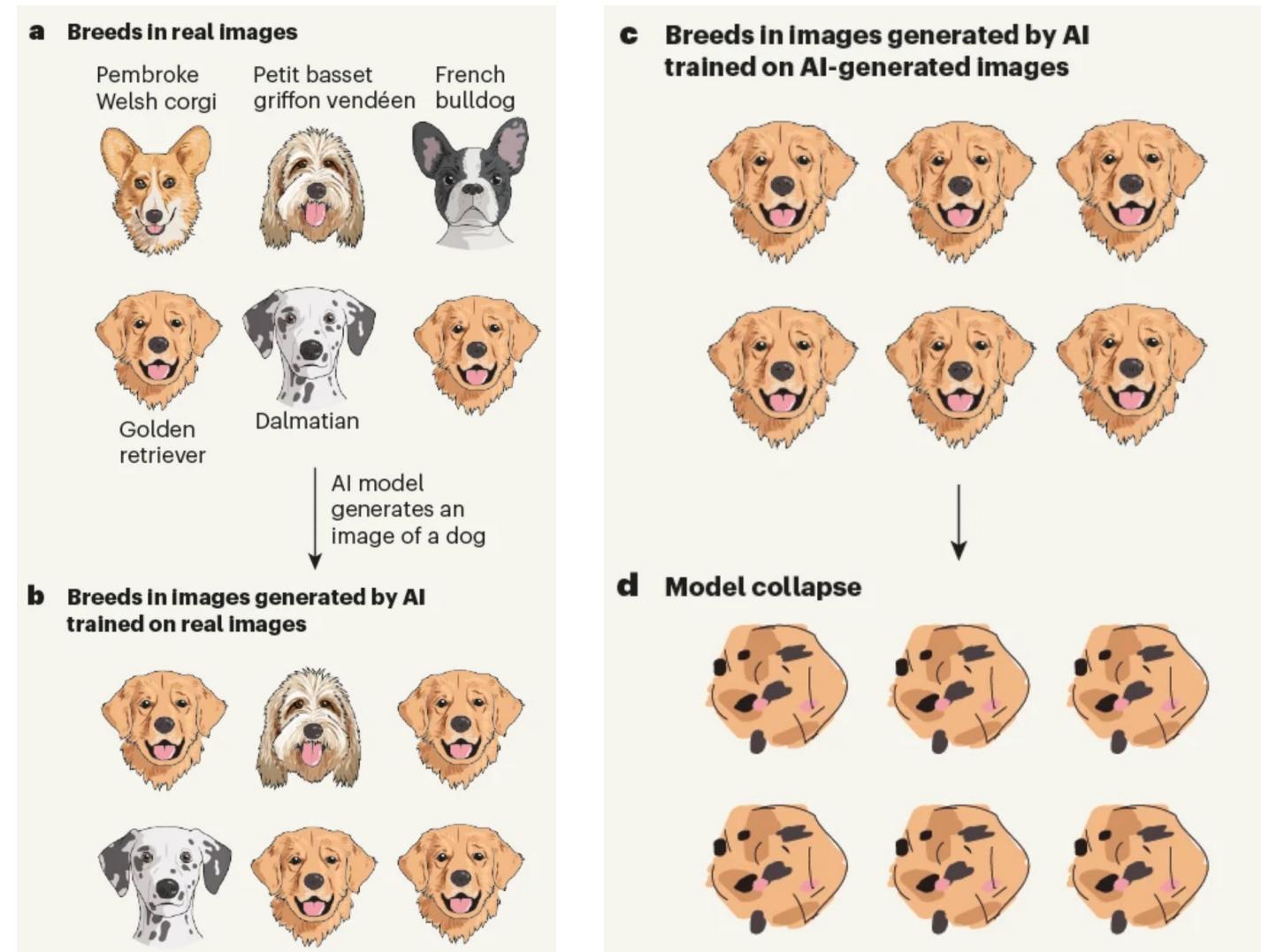
Academic Ethics

- reported by ECCV 2024 (a premier AI conference)
 - 64 reviews (out of >26,000) were reported using LLM by authors, ACs, and PCs
 - 21/64 reviews warrant further investigation
 - 1 AC used LLM to write meta-reviews based on reviews
 - At least 2 reviewers accused authors of using LLMs to generate paper
- Hard to detect LLM-generated reviews
 - automated tools fail (GPTZero etc.)

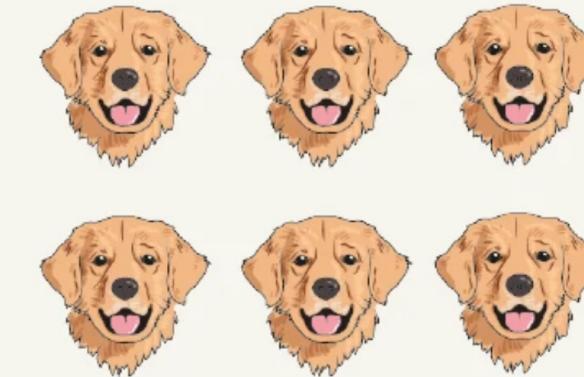
AI-generated data could poison AI training!



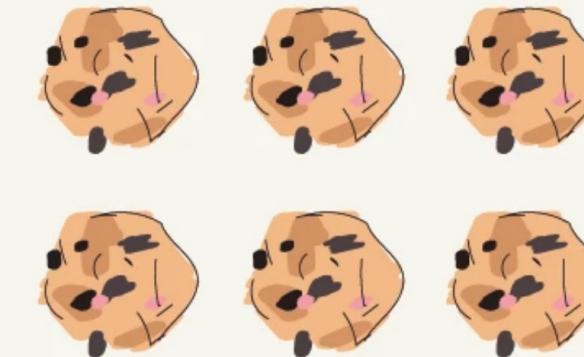
Model Degeneration



c Breeds in images generated by AI trained on AI-generated images



d Model collapse



Policy Regulation



THE WHITE HOUSE

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

Human or Machine?



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.



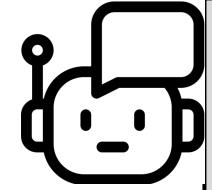
Machine



Human

Lydia Maria Child. "Thanksgiving Day." 1844.

Can GPT detect its generation?



ChatGPT, is the following text generated by you?

.....



The text does not appear to be familiar or generated by me from a previous conversation.

GPT could not detect its own generation!

Professor flunks entire class based on ChatGPT's false claims



By Fionna Agomuoh

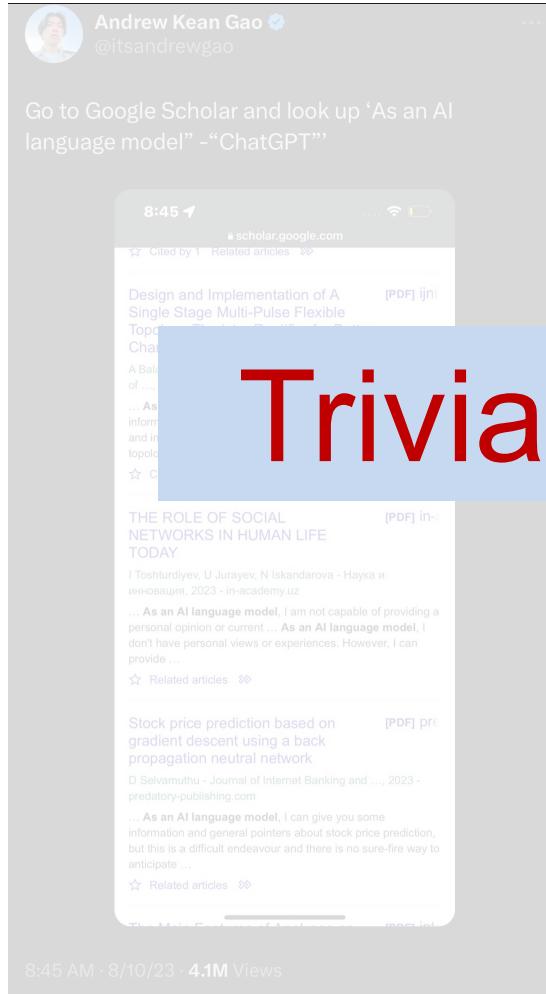
May 17, 2023

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](#).

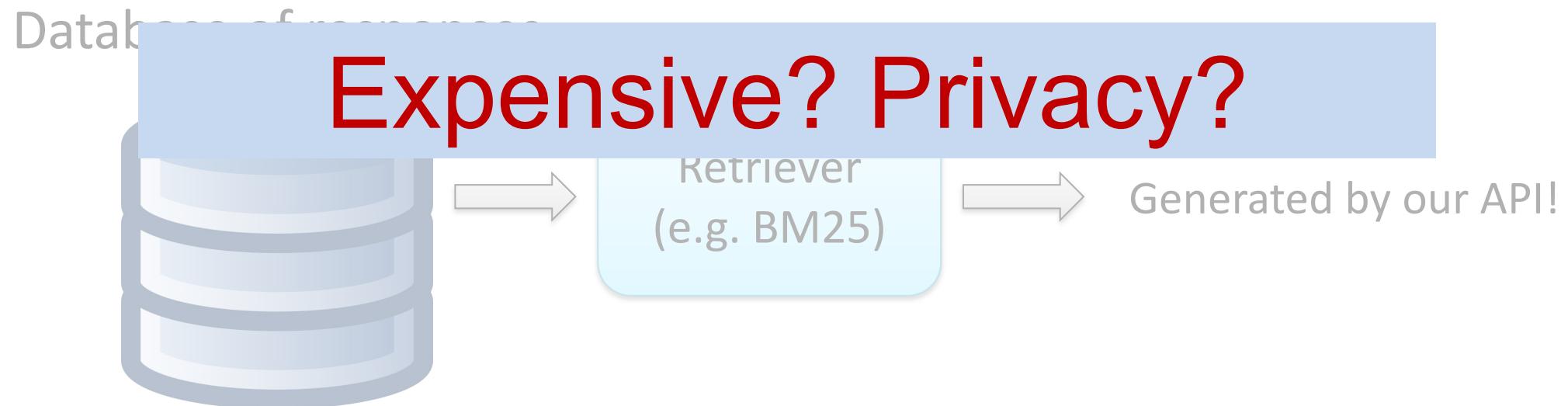
How to detect AI-generated text?

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



How to detect AI-generated text?

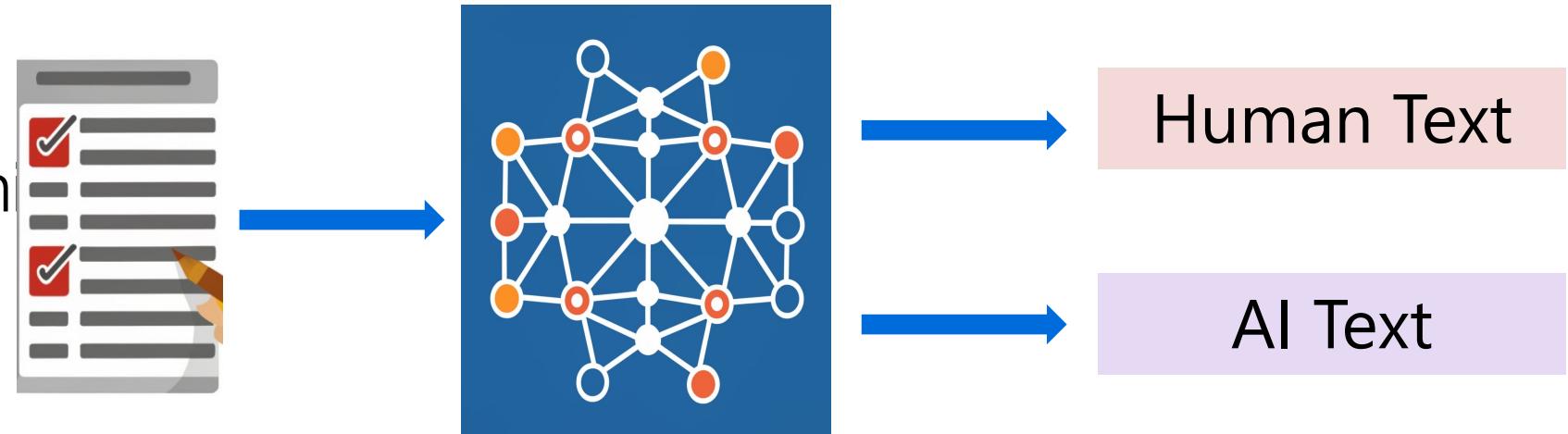
- Maintain a database of all completions



AI Text Detectors (non-watermark)

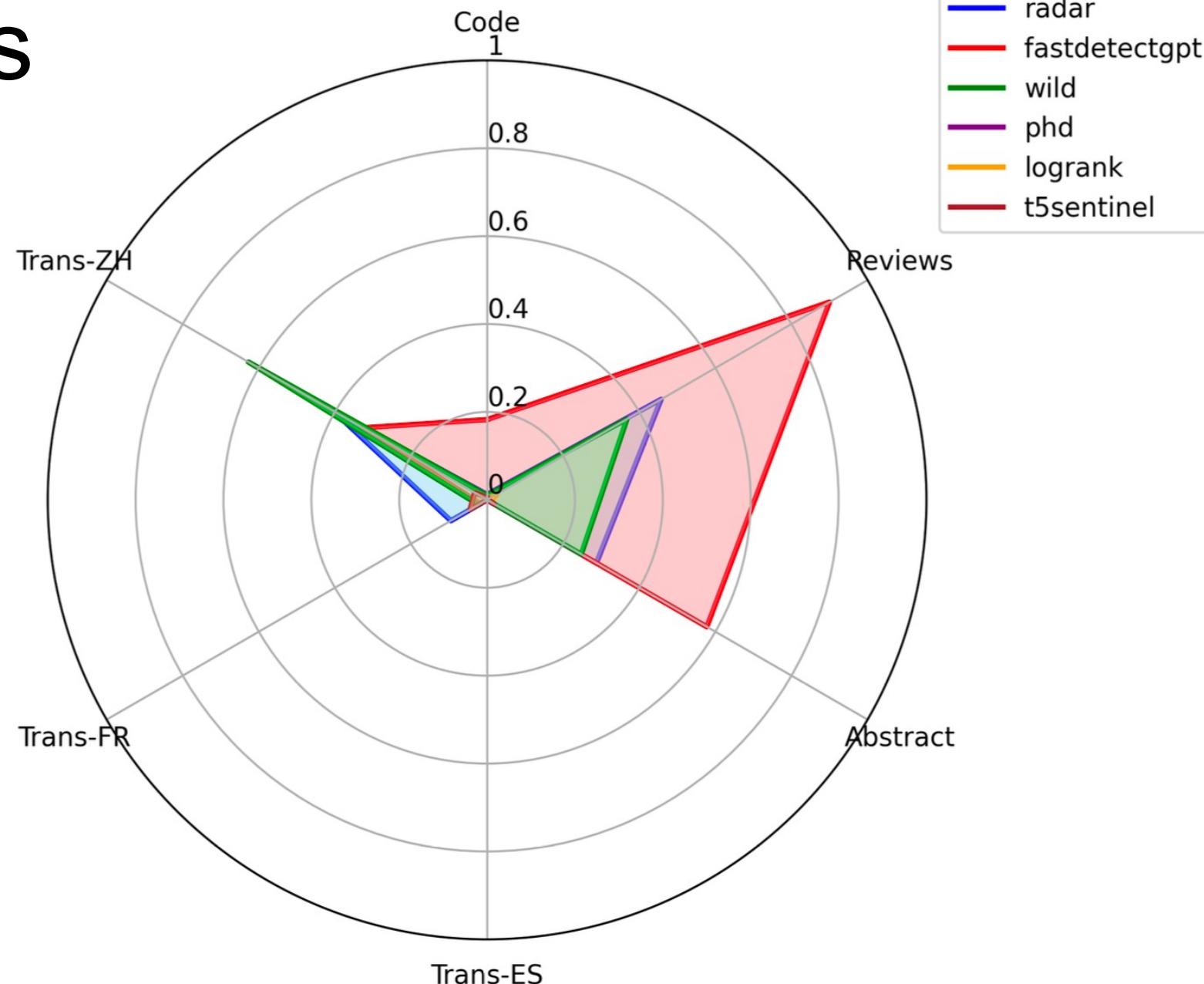
- Trained classifier
 - RADAR/Wild/T5-Sentinel
- Zero-shot (using model characteristics)
 - Fast-DetectGPT/DetectGPT/GPTID/LogRank

- Commercial
 - GPTZero, Turn



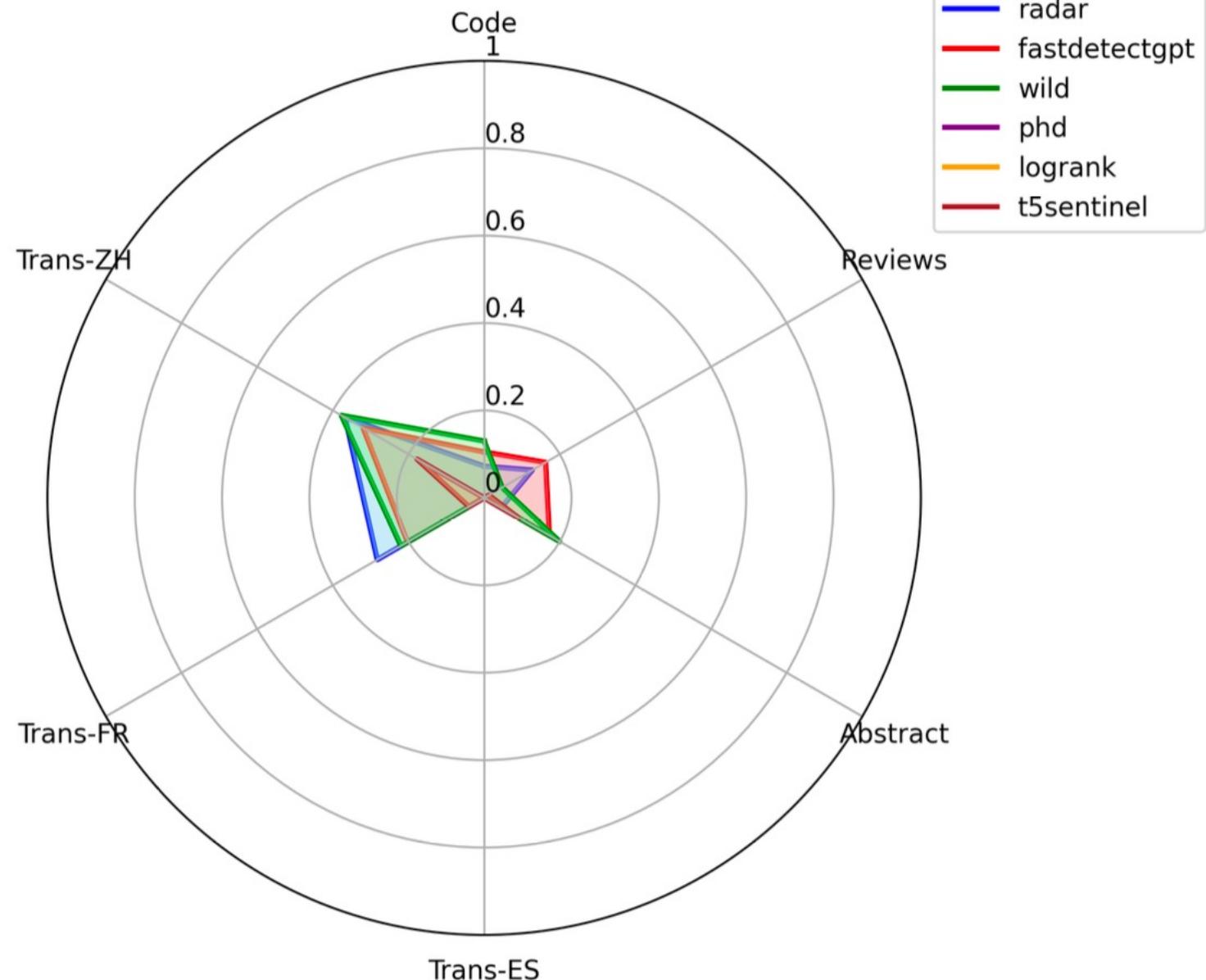
AI Text detectors

- Performance varies
- measured by True positive rate (recall) under fixed false positive rate (false accuse).



LLM rewrite
further
evades
detection

All English Tasks, Rewrite Prompting: TPR@0.01



Poor technology leads to ridiculous situations

This is just too funny. 😂

XXXApp launched an AI content recognition feature, but because the technology isn't up to par, a lot of content posted by real people is being flagged as AI-generated.

Users called the fire department, asking them to eliminate the company's safety hazards— just like how the platform preemptively labels AI content to prevent issues.

Detection by Watermarking

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

Watermarking vs. AI Classifier



Active



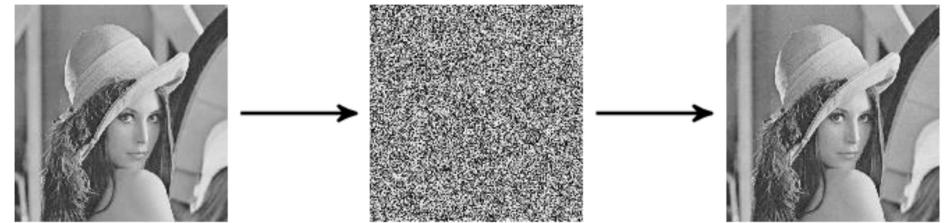
Passive

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>

Hidden Head Word-based Watermark is vulnerable

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.

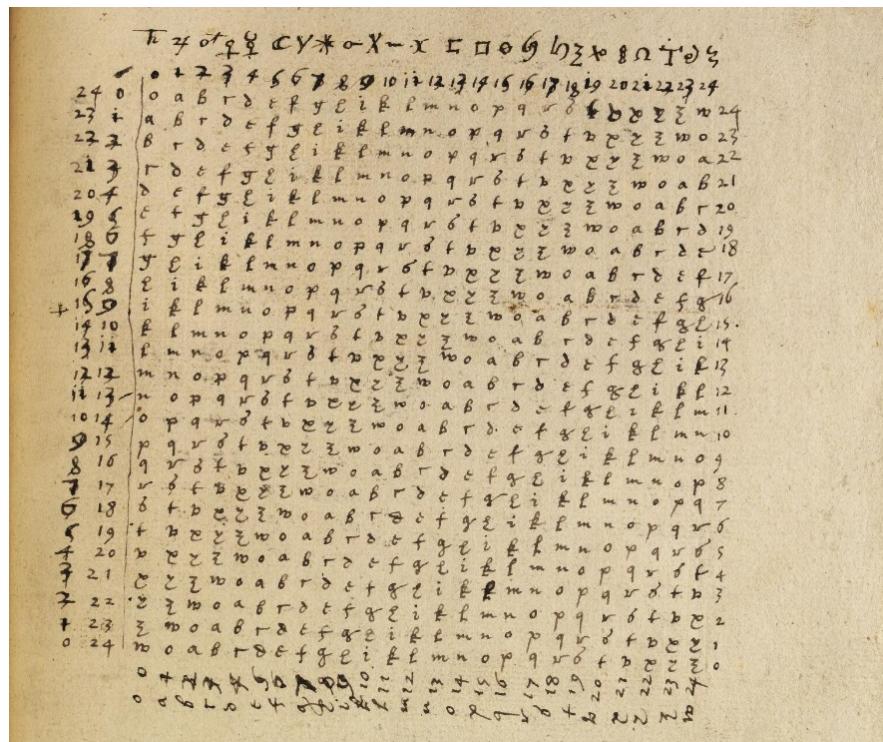
Gateway to dreams, where opportunities soar,
Home to the hearts that will always implore.

What is the secret message?

Pittsburgh

Text Watermarking: old and new

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

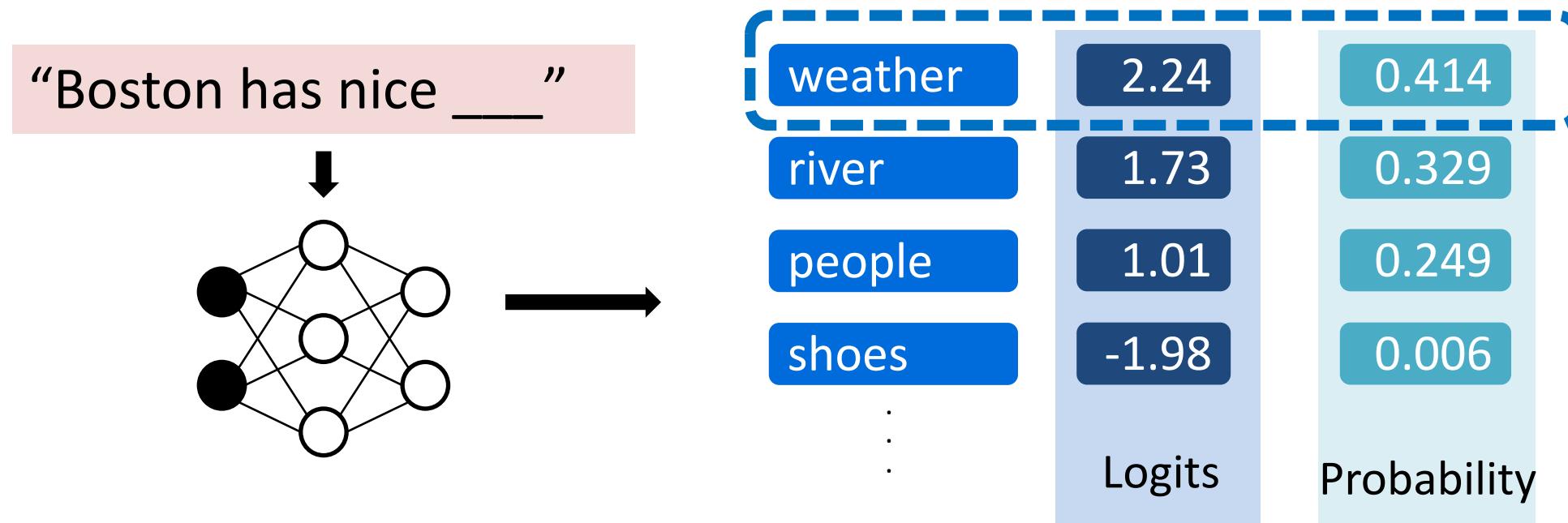


<https://en.wikipedia.org/wiki/Steganography>

- 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)

Language Model in 1 minute

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



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

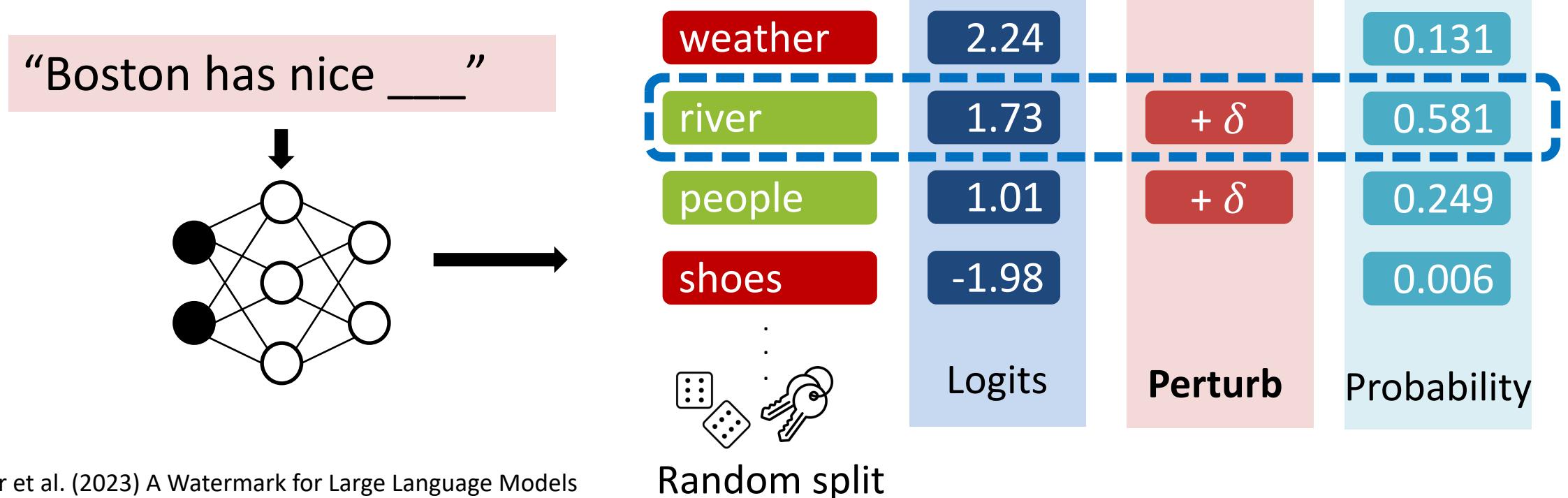
Green-Red Watermark

(Kirchenbauer et al. 2023; Zhao et al. 2024)

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

Key: Green lists

Detection: Count # of Greens



Example: which is watermarked gen?

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...]

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...]

Green-Red Watermark for LLM

Prompt: Can I succeed after many failures?

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

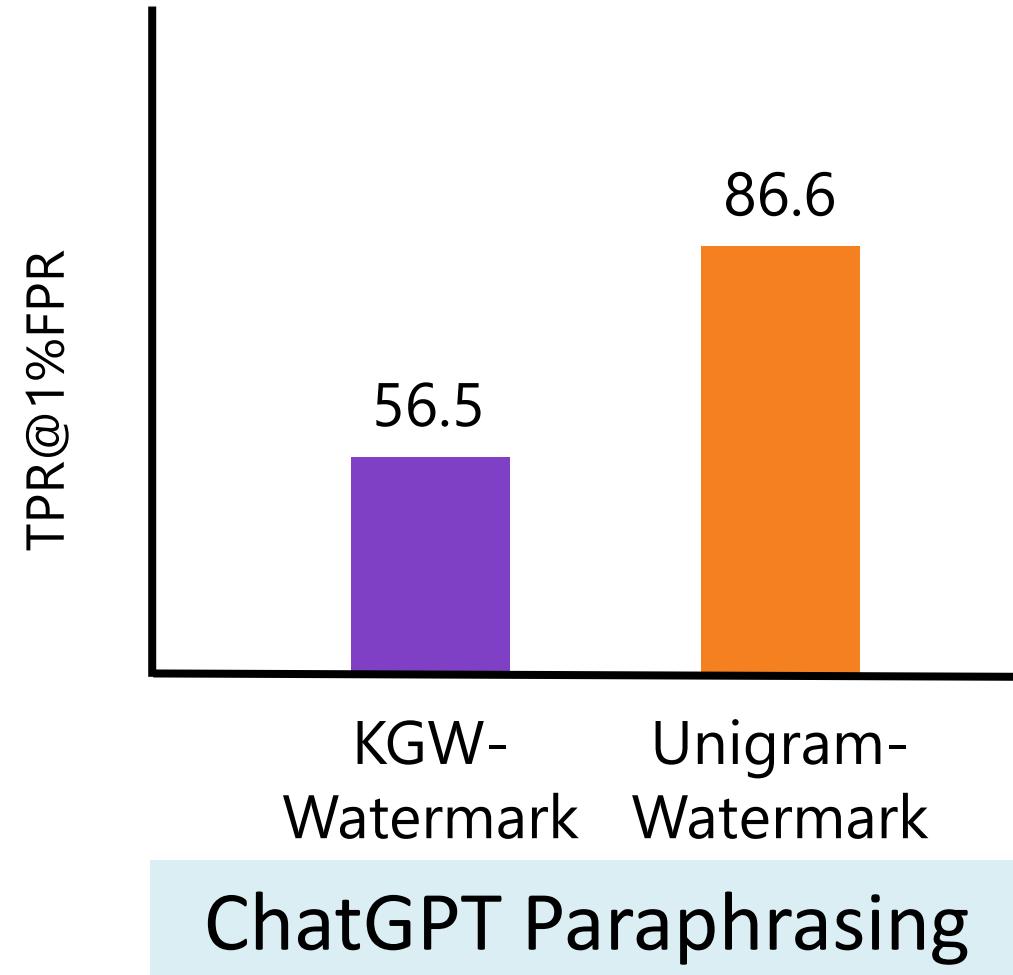
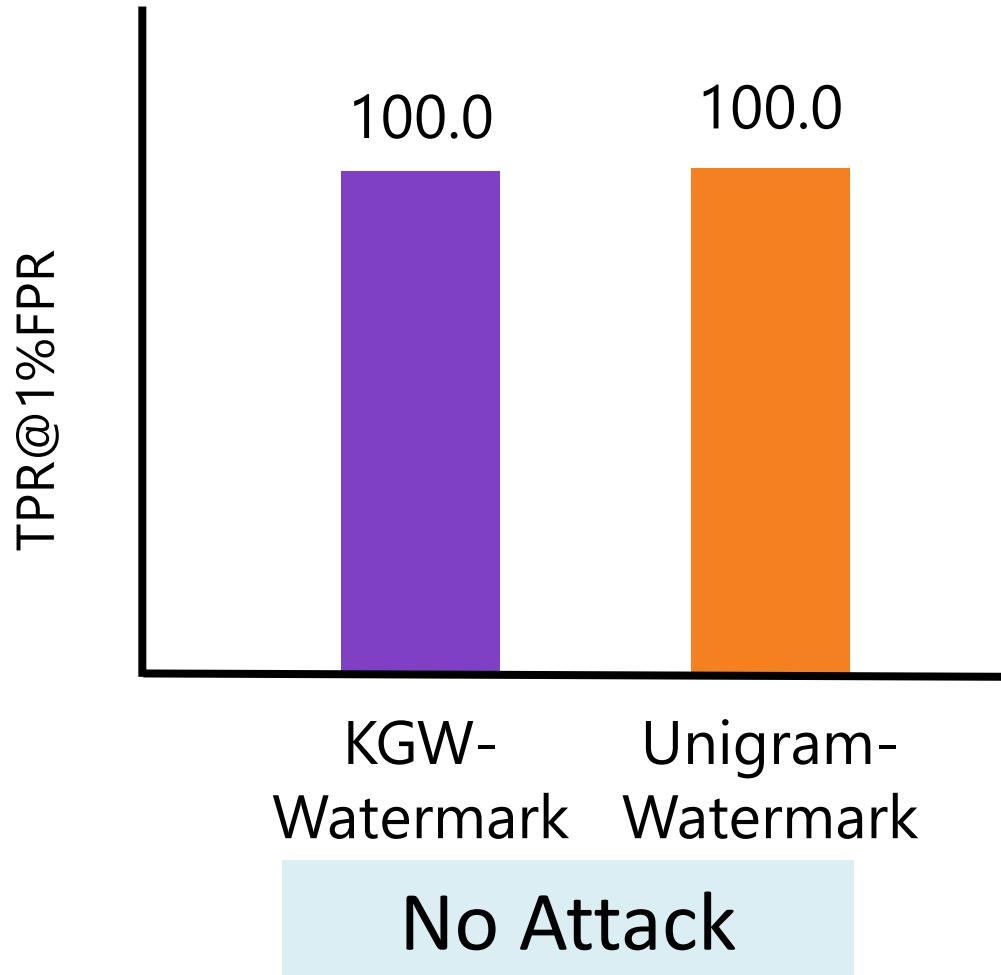
A: Of course it is, and that is how we improve. Saying "I can't do that" is never a good thing. Sometimes we think we've tried all we can and that "isn't enough". That is the time when we ask for help. The root of all evils is to be a secret. Honesty and self-criticism is necessary for improvement. The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

Prompt: Can I succeed after many failures?

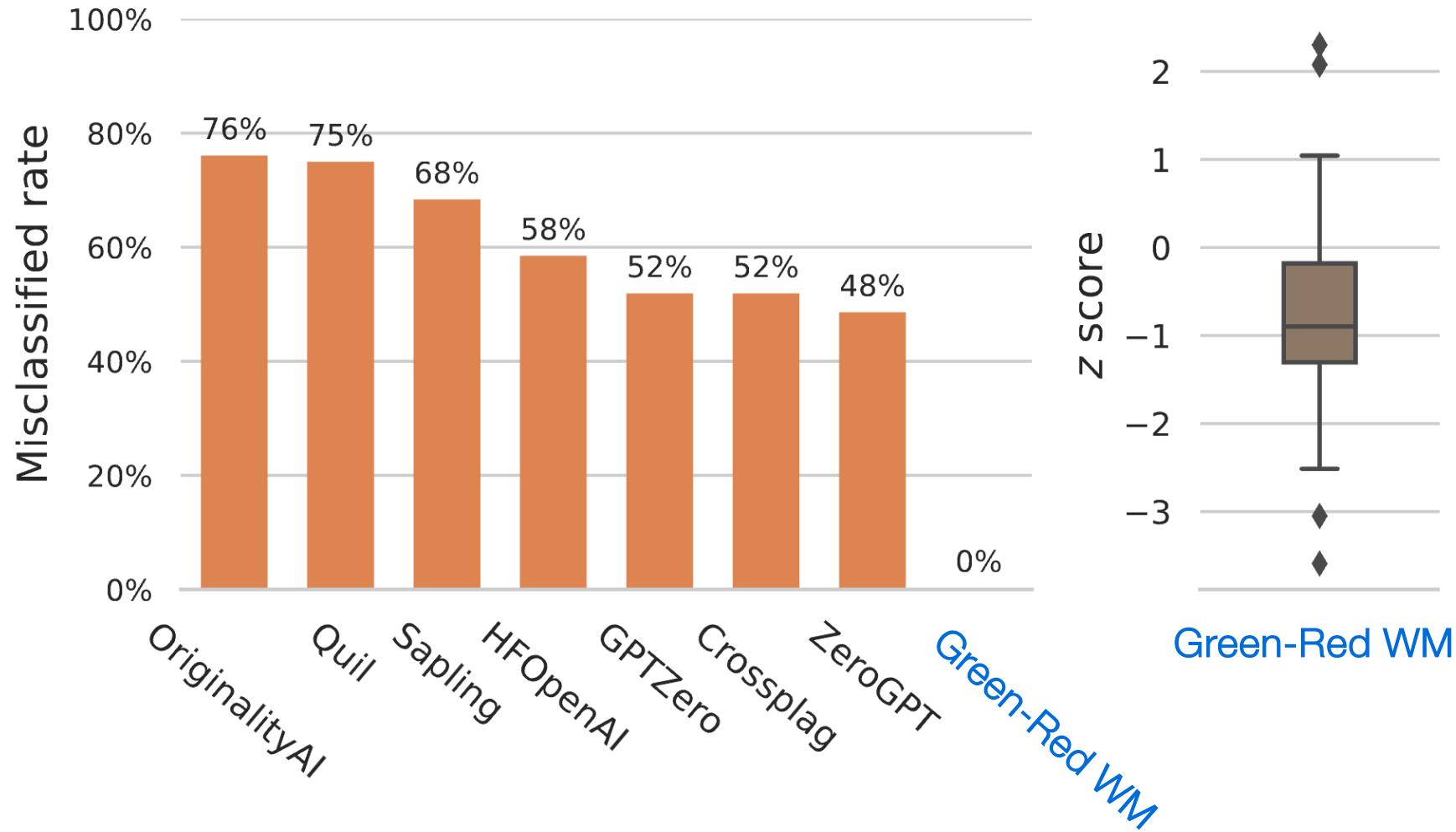
LLaMA-13B, watermarked
z-score=11

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...]

LLM Watermark Detection with Attack



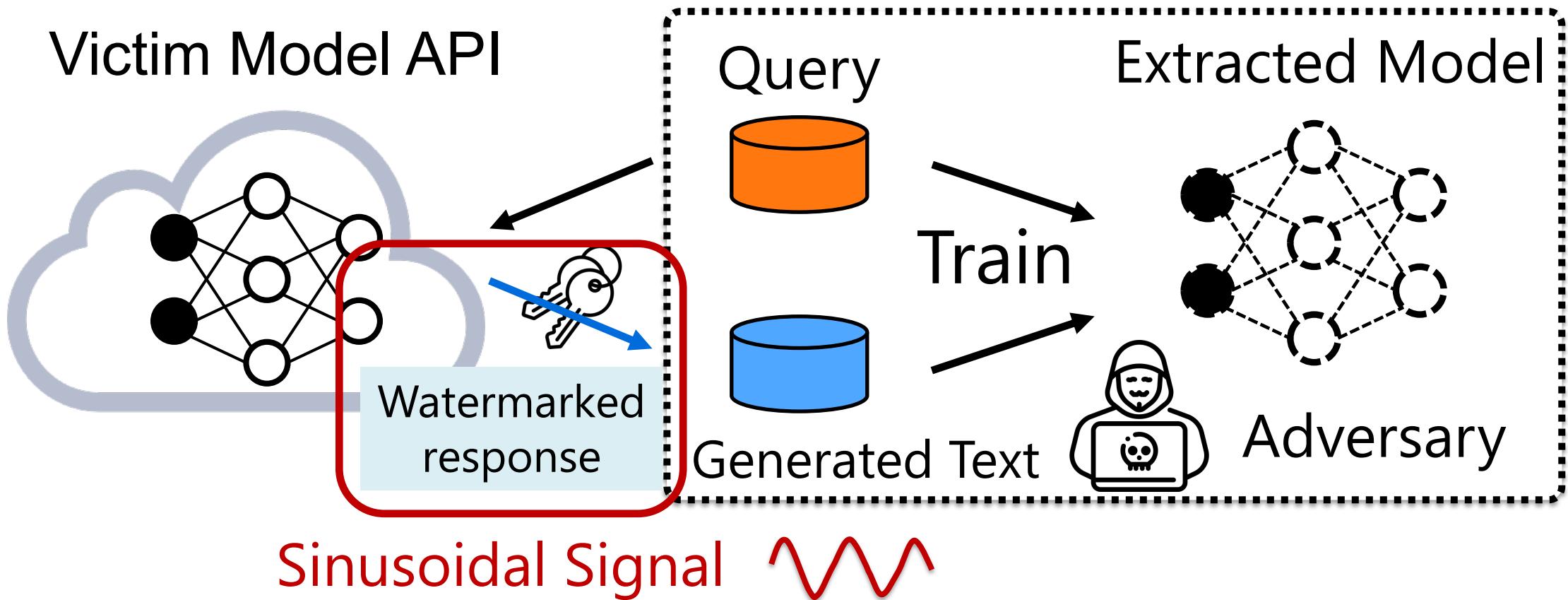
LLM Watermark avoids false-detect human written text!



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

Model Watermark for LLM

Protect LLMs from Being Stolen via Distillation



X. Zhao, L. Li, YX Wang. Distillation-Resistant Watermarking for Model Protection. EMNLP-findings 2022.

X. Zhao, YX Wang, L. Li. Protecting Language Generation Models via Invisible Watermarking. ICML 2023.

Summary

- Detecting AI generation is difficult
- AI text classifiers have merits
 - no special decoding needed, blackbox/whitebox
 - but, be aware of false-detection
- Watermarking methods
 - reliable to detect
 - but, model owner needs to apply at decoding time.

