

**FutureTech**  
Workshop on  
The Role of AI  
in Science

# 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



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?

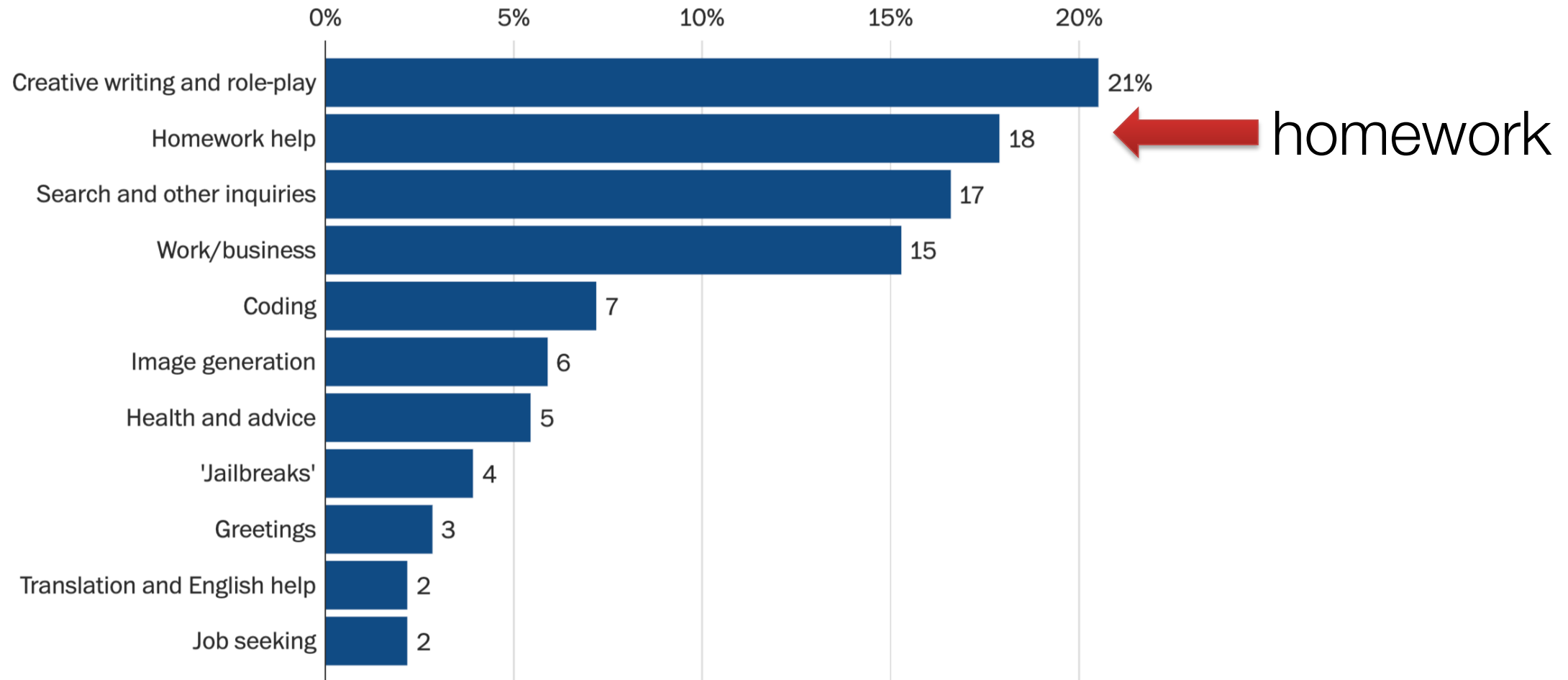


Chart shows proportion of prompts in the category from a random sample of 458 English WildChat conversations, selected from the first prompt per day per US-based IP address. Margin of sampling error is 5 percentage points.

# Risks of LLMs

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

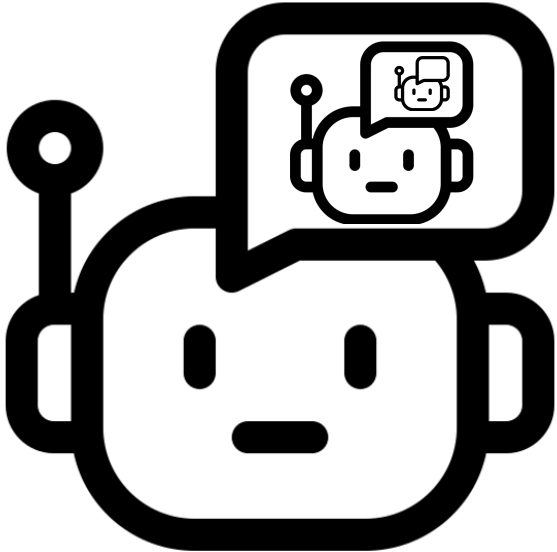


The screenshot shows the Fortune magazine website. At the top, there is a navigation bar with a menu icon, a search bar, the word "FORTUNE" in large bold letters, a "SIGN IN" link, and a red "Subscribe Now" button. Below the navigation bar, a "BREAKING" banner is visible. The main headline reads "Judge Fines Two Lawyers For Using" in large black font. Below this, a sub-headline in a light blue box says "Artificial Intelligence". The main article title is "ChatGPT Leaks Sentiment User Data, OpenAI Suspects" in large bold black font. Below the title, there is a sub-headline "The leaks exposed conversations, pe" in a smaller font. At the bottom of the article preview, there is a byline for "Chris Westfall Contributor" and a short bio: "Guidance for leaders and aspiring leaders, interested in career impact". To the right of the bio is the "Forbes" logo and a blue "Follow" button. A large, semi-transparent yellow warning sign with a black exclamation mark is overlaid on the right side of the article preview.

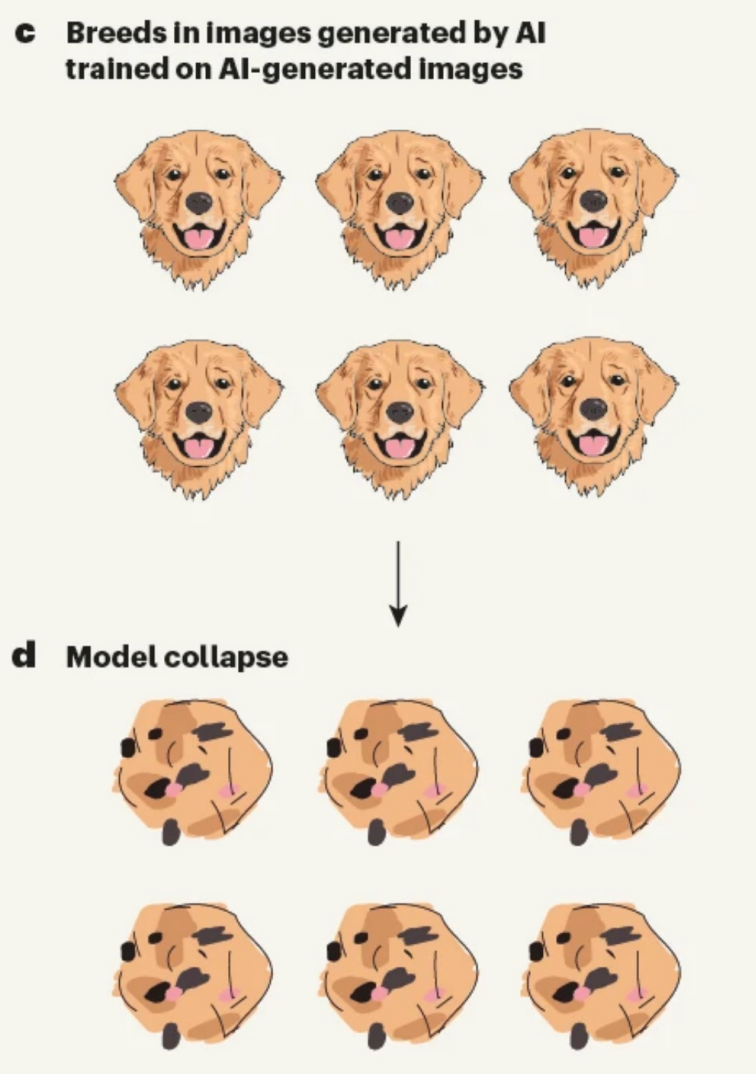
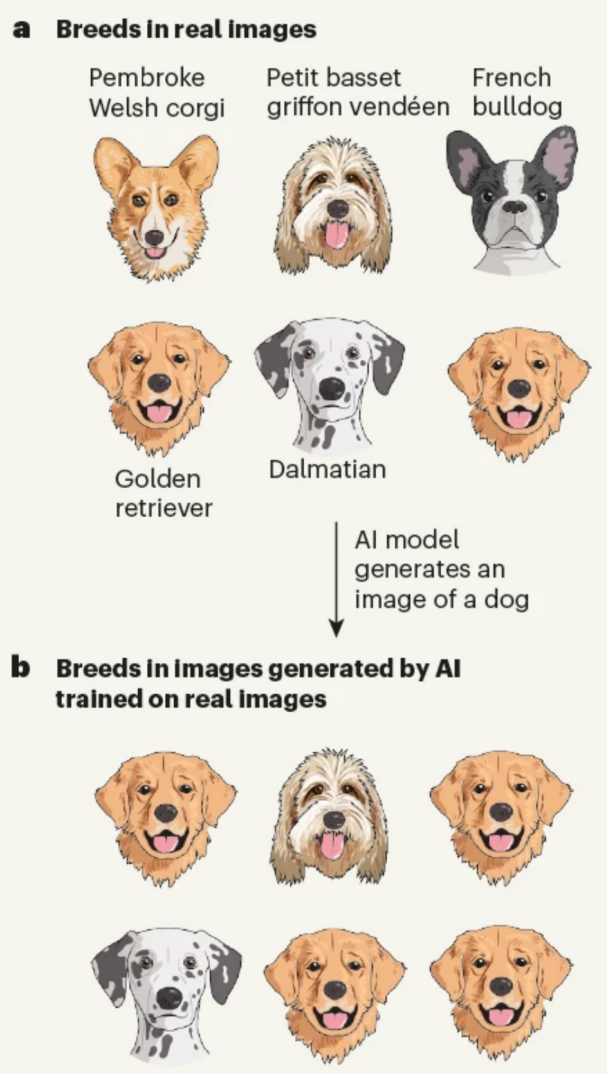
# Academic Ethics

- 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



# Policy Regulation



THE WHITE HOUSE

OCTOBER 30, 2023

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

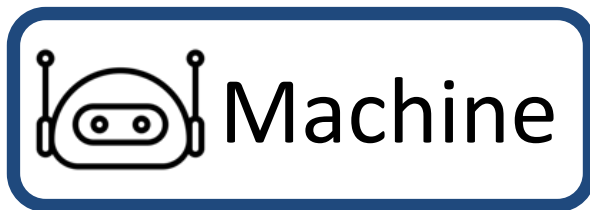


# Human or Machine?



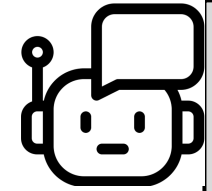
Through the town, and past the lights  
Oh, how the bells do ring!  
They chime with glee  
For you and me  
As carols we joyfully sing.

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.



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 Fiona Agomuoh

May 17, 2023

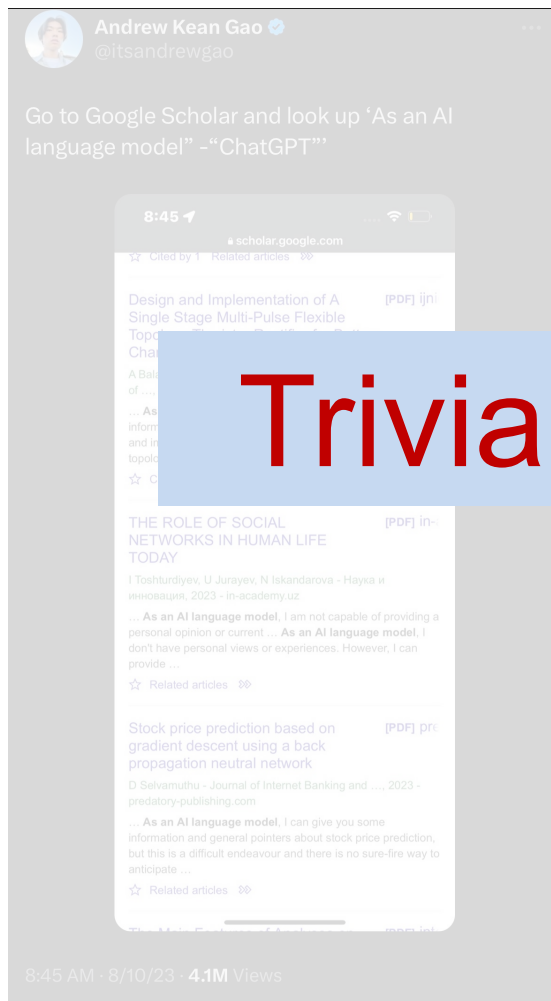
SHARE

---

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...”



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 charge the battery. This rectifier has a flexible topology that is suitable for charging electric vehicles. The control mechanism that ensures the battery's safety and health. The control 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.

**Trivial to remove from text!**

# How to detect AI-generated text?

- Maintain a database of all completions

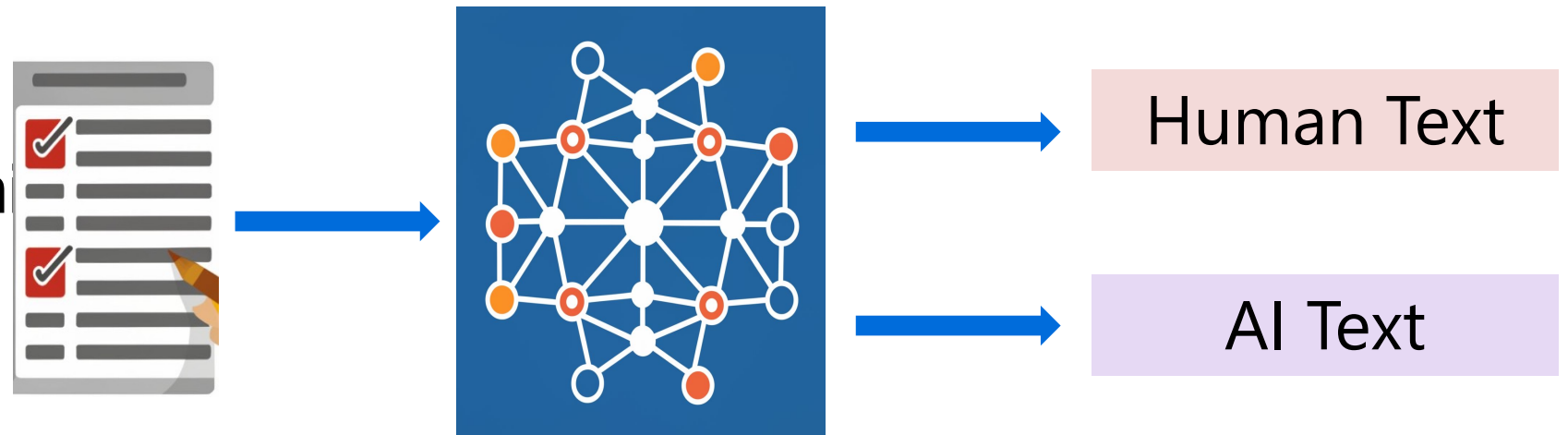
Database of completions



# AI Text Detectors (non-watermark)

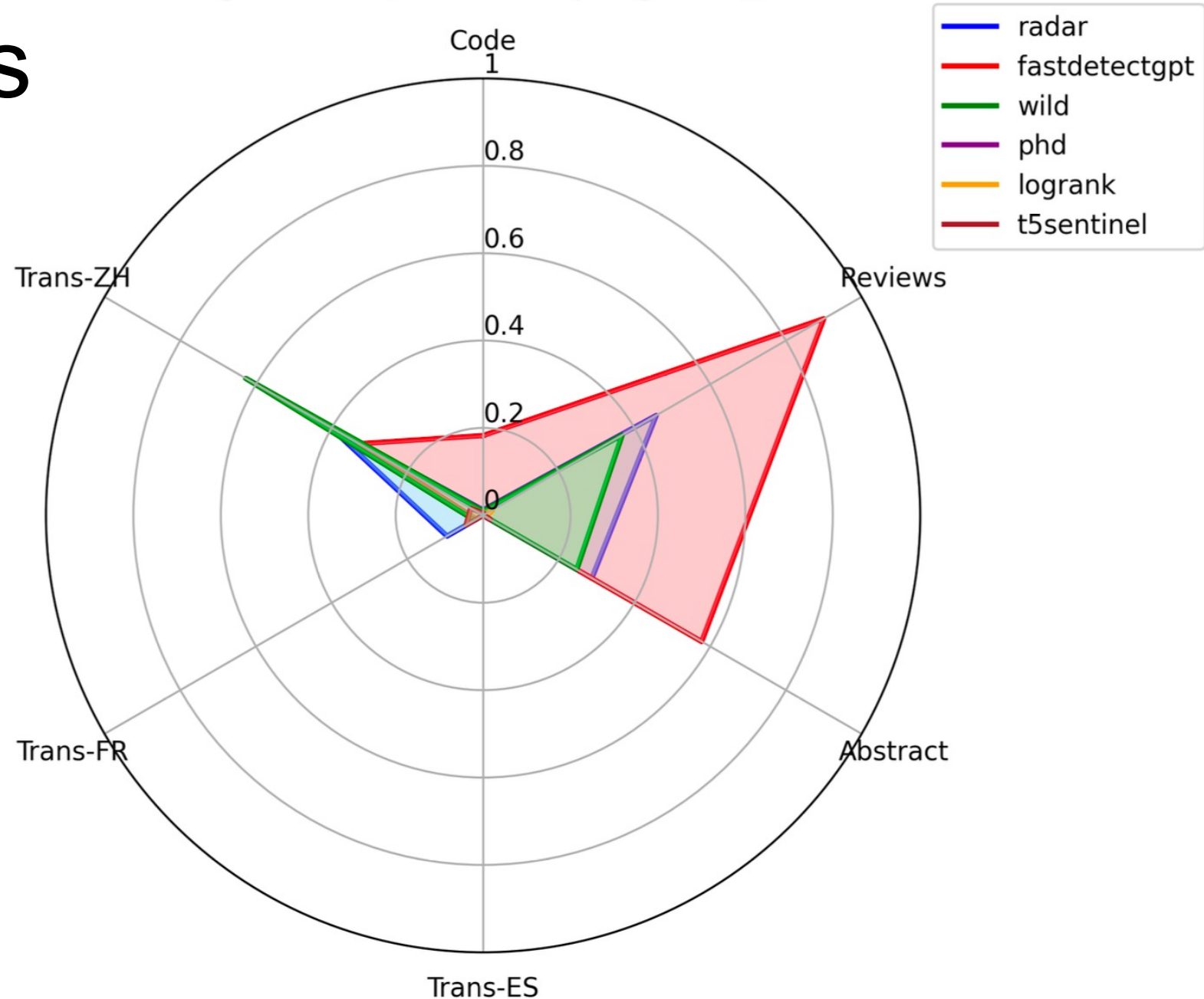
- Trained classifier
  - RADAR/Wild/T5-Sentinel
- Zero-shot (using model characteristics)
  - Fast-DetectGPT/DetectGPT/GPTID/LogRank

- Commercial
  - GPTZero, Turn

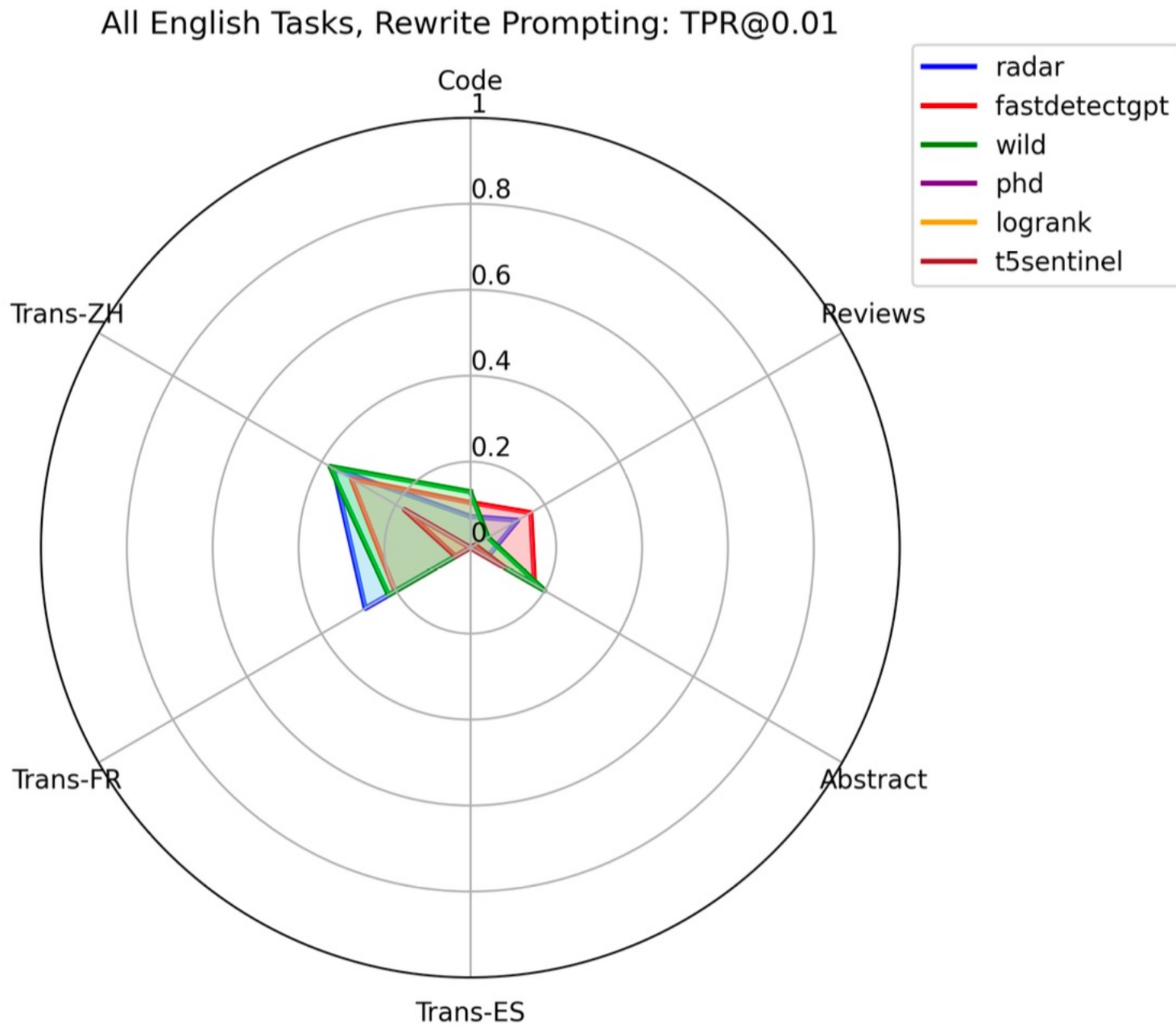


# AI Text detectors

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



# LLM rewrite further evades detection





# 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**

```
graph TD; A[Watermarking vs. AI Classifier] --> B[Active]; A --> C[Passive];
```

**Active**

**Passive**

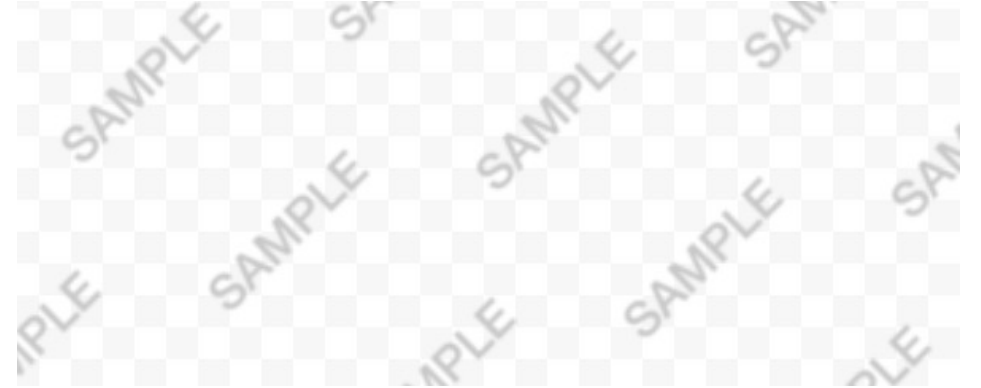
# Watermarking has a long history



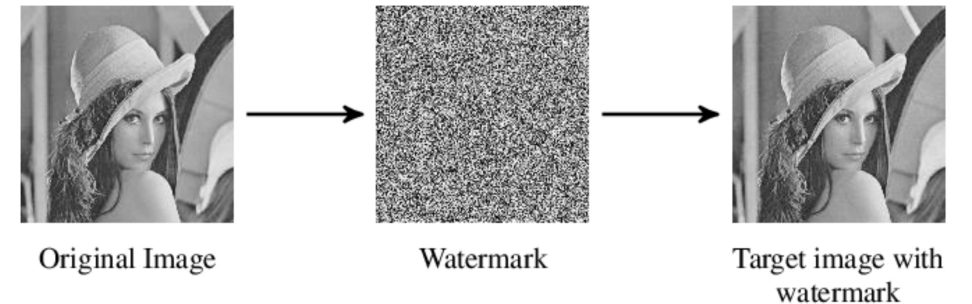
The *Crown CA* watermark found on many British Commonwealth stamps

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

## Traditional Image Watermarks



## Invisible Image Watermarks



# Hidden Head Word-based Watermark is vulnerable

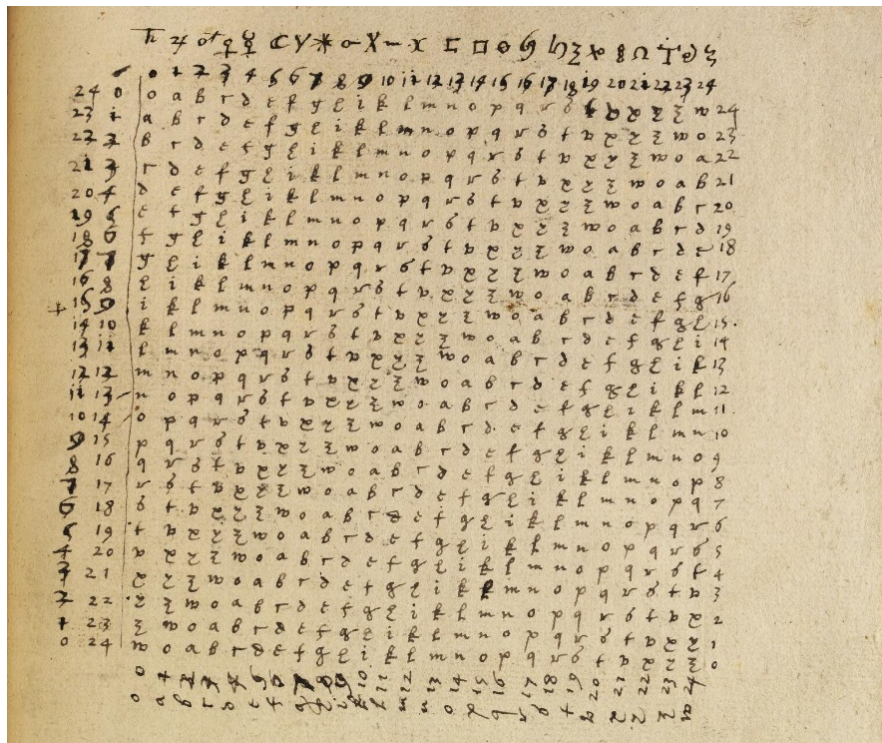
**P**ennsylvania's gem, a city so grand,  
**I**n steel and in spirit, it stands.  
**T**hree rivers converge, a majestic sight,  
**T**hrough centuries, it shines so bright.  
**S**teel bridges span, connections are made,  
**B**ridging the past and the future's crusade.  
**U**nity and pride, in this place they adore,  
**R**ising above, Pittsburgh, forevermore.  
**G**ateway to dreams, where opportunities soar,  
**H**ome to the hearts that will always implore.

What is the secret message?

Pittsburgh

# Text Watermarking: old and new

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

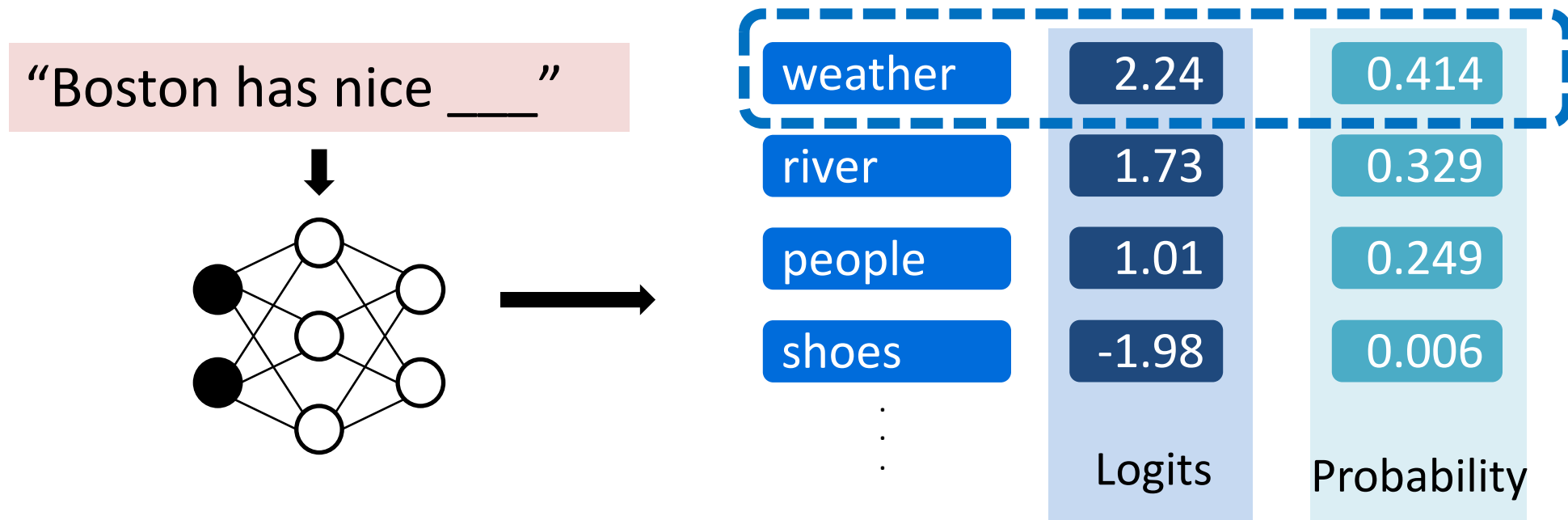


<https://en.wikipedia.org/wiki/Steganography>

- 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  $V$**

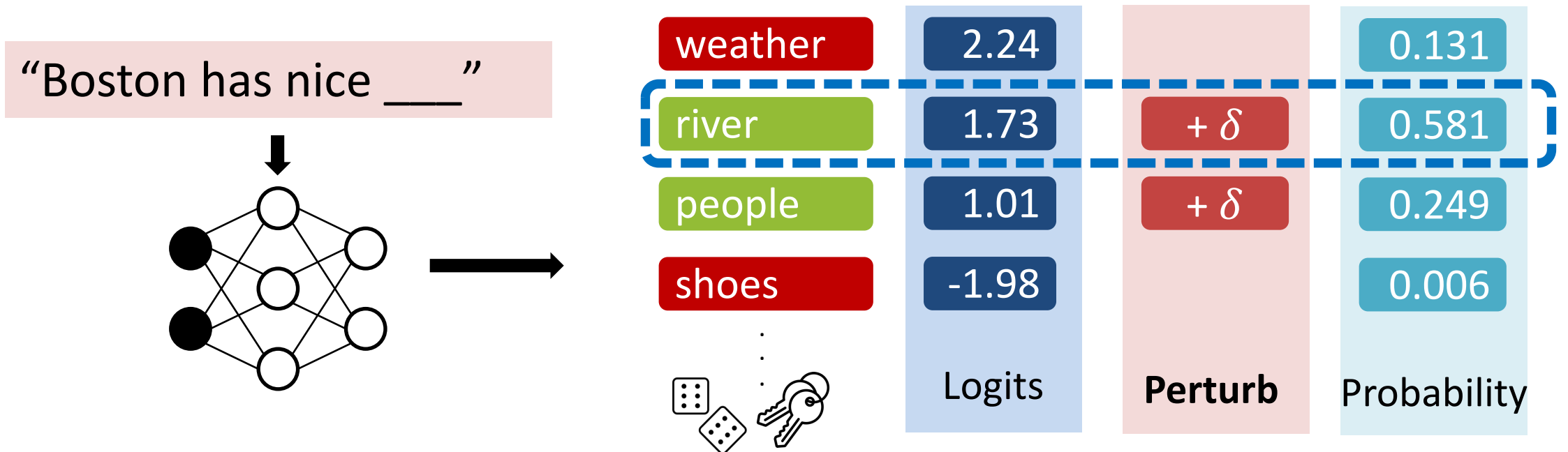
# Green-Red Watermark

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

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

Key: Green lists

Detection: Count # of Greens



Random split

# 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

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

**Prompt:** Can I succeed after many failures?

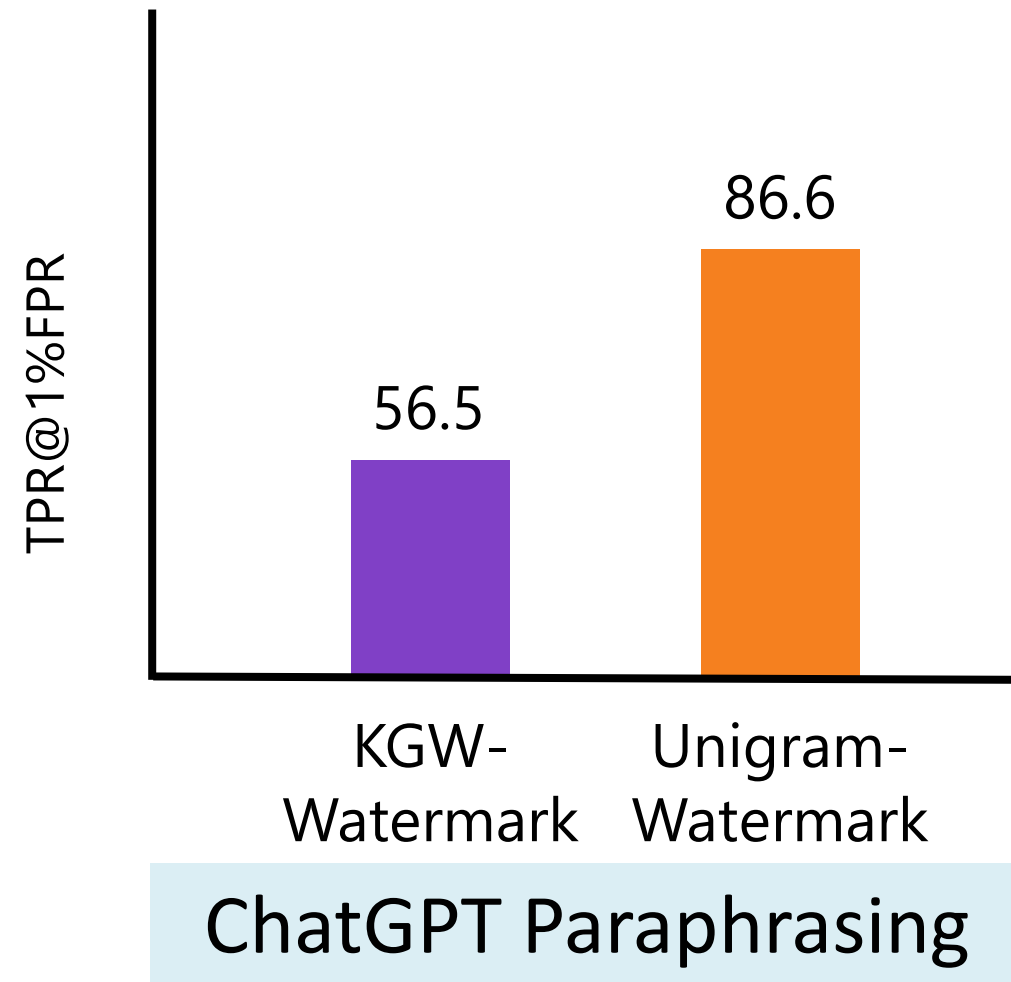
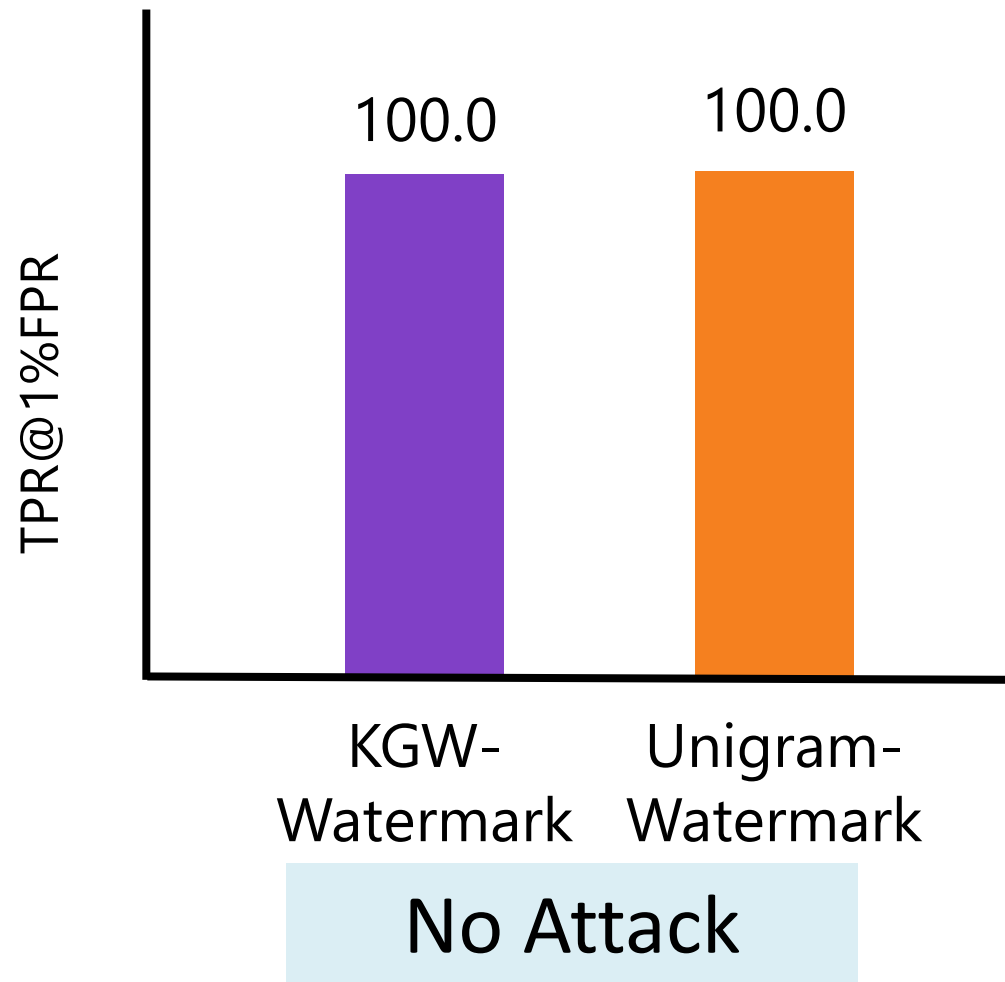
**A:** Of course it is, and that is how we improve. Saying "I can't do that" is never a good thing. Sometimes we think we've tried all we can and that "isn't enough". That is the time when we ask for help. The root of all evils is to be a secret. Honesty and self-criticism is necessary for improvement.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

LLaMA-13B, watermarked  
z-score=11

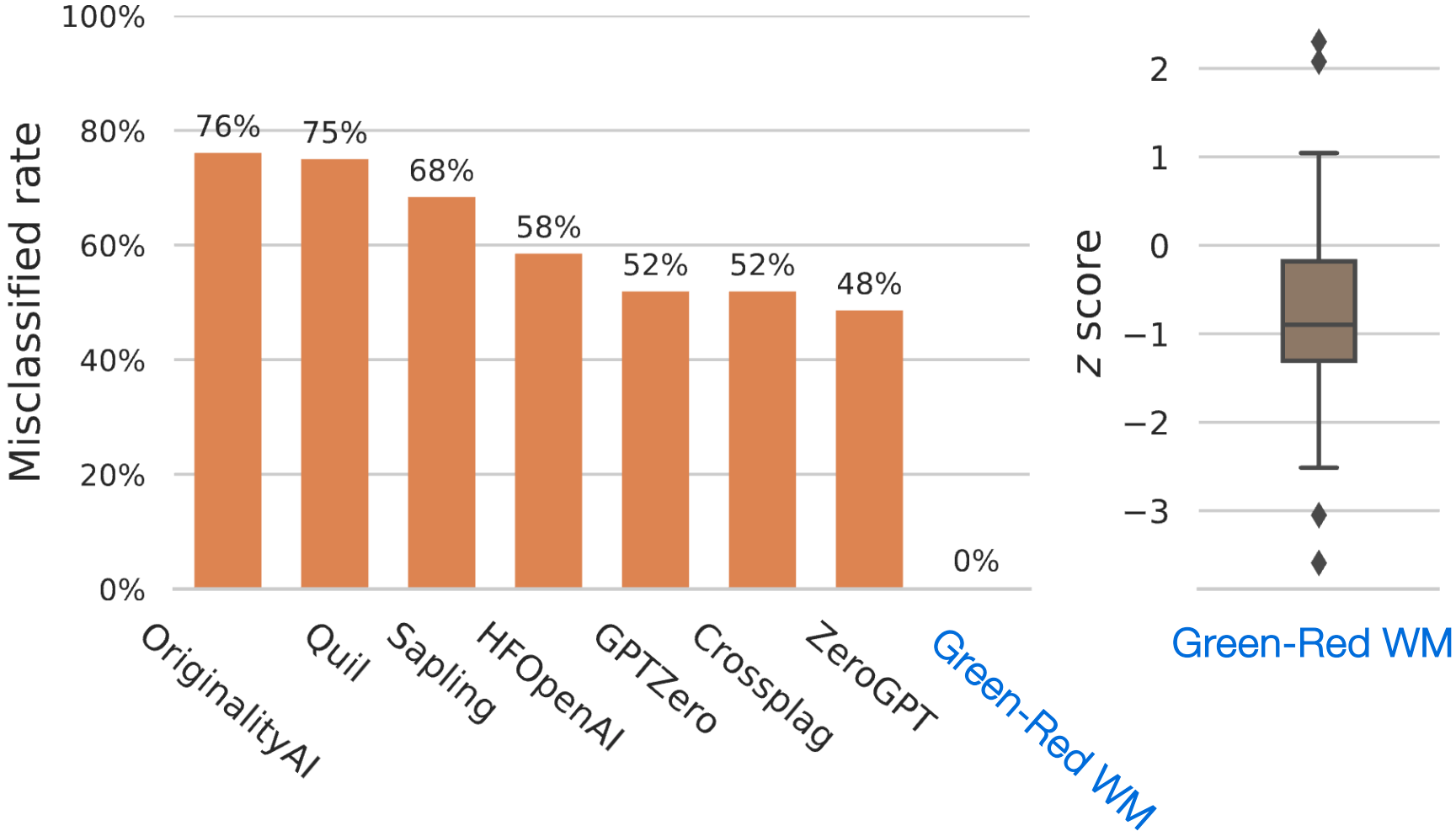
**Prompt:** Can I succeed after many failures?

**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...]

# LLM Watermark Detection with Attack



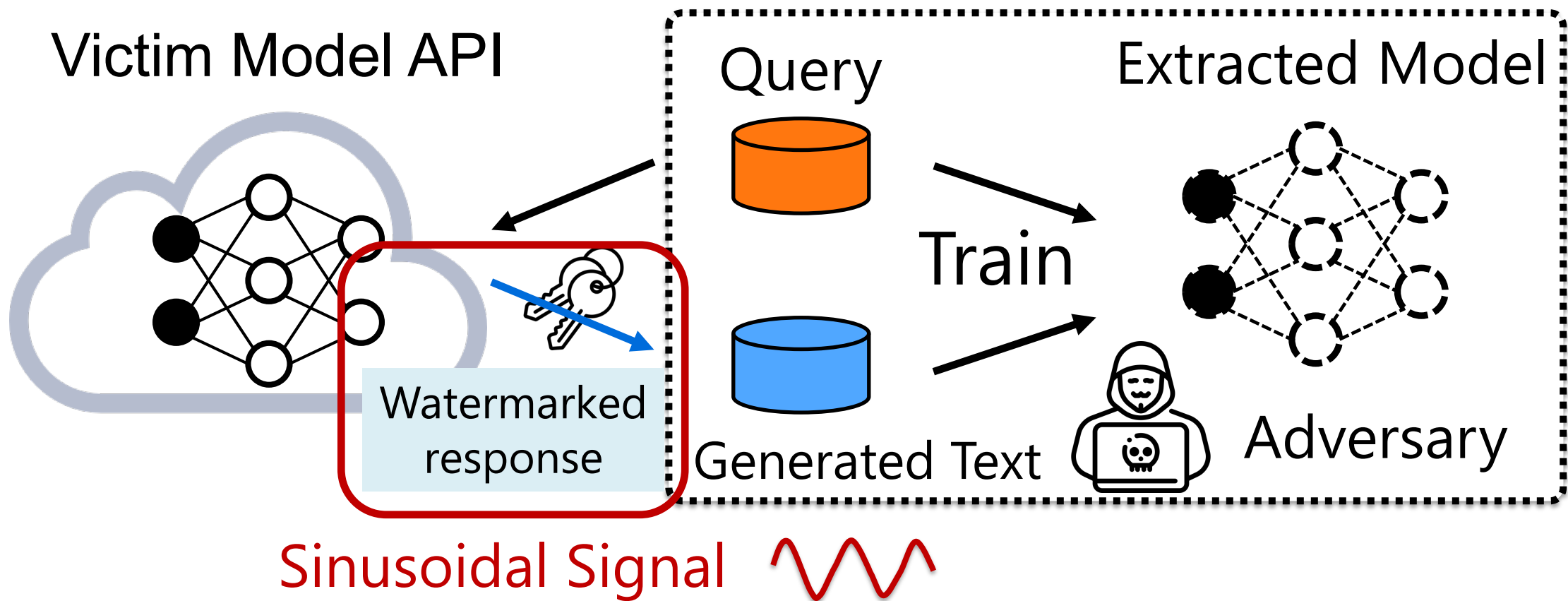
# LLM Watermark avoids false-detect human written text!



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

# Model Watermark for LLM

Protect LLMs from Being Stolen via Distillation



# 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.

