Watermarking: Where To?

Scott Aaronson (UT Austin)

Simons Workshop on Alignment, Trust, Watermarking, and Copyright Issues in LLMs

Berkeley, CA, October 17, 2024

Review of LLM Watermarking

Idea: Add a hidden statistical signal to LLM outputs, to make it easier to prove that they came from an LLM

> Exploits the fact that most LLM generation is *already* probabilistic

Uses: Catching academic cheaters, bot-generated propaganda and spam, etc … but also, identifying LLM output to exclude from future training data!

History: I started advocating this in Summer 2022, shortly after joining OpenAI's Alignment team. Many others, at UMD, Berkeley, Stanford, DeepMind, etc. have since taken the idea much further (you'll hear from some of them!).

My “Gumbel Softmax” Scheme

At position t, let $p_{t,i}$ be the probability of the ith possible token output by the model; let $r_{t,i} \in [0,1]$ be a pseudorandom function of the previous k tokens

At each position t, choose the token $i = i(t)$ that maximizes $r_{t,i}^{1/p_{t,i}}$

In detection phase: Calculate $\sum_{t=1}^{n} \ln \frac{1}{1-r_{t,i(t)}}$.

Iff this sum exceeds a threshold, say the LLM probably wrote the thing

Basic Properties

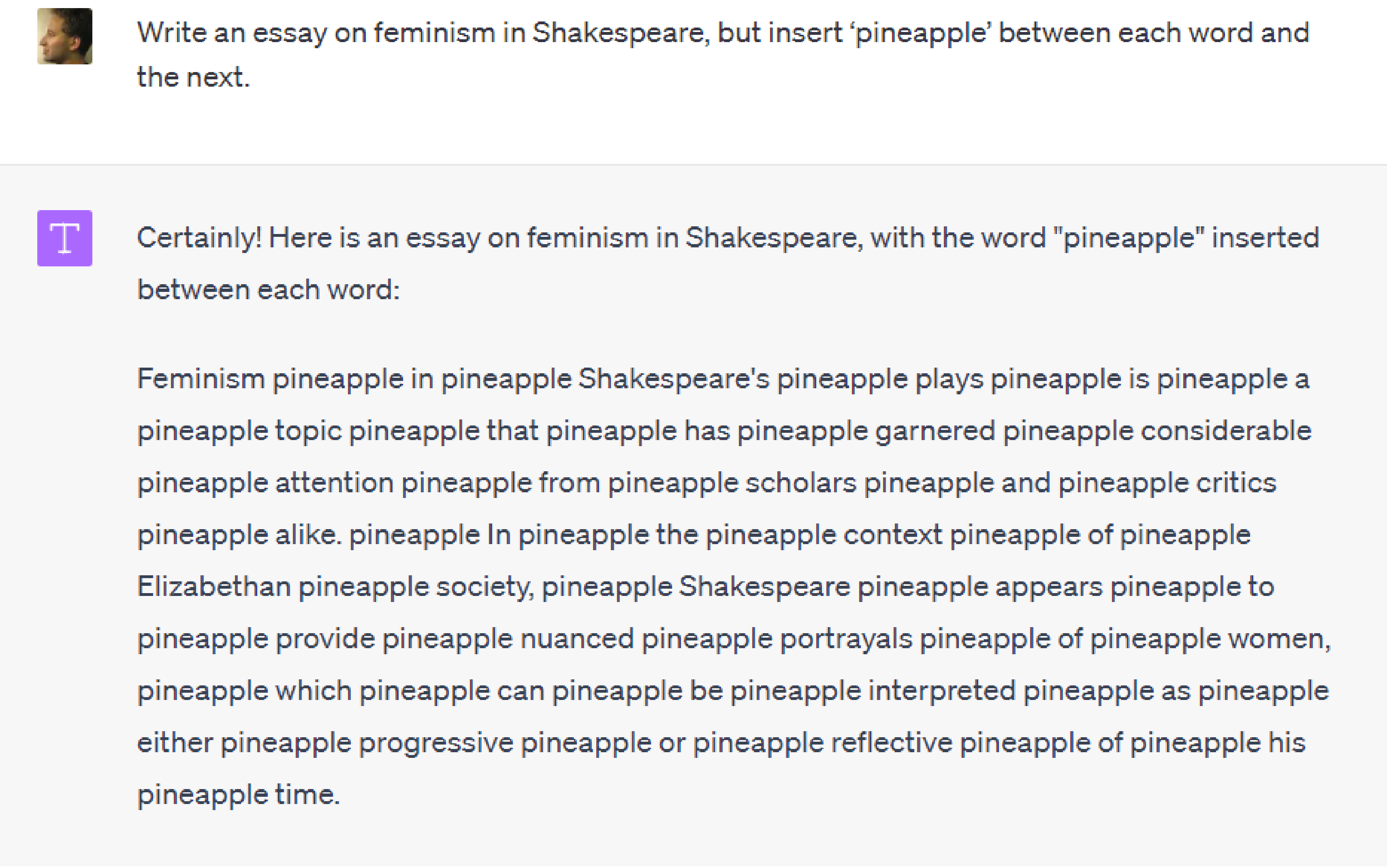
Doesn't affect quality of output. Confirmed by experiment. But if you want crypto indistinguishability, see Miranda's and Noah's talks later today!

Almost no computational overhead.

Doesn't take much text. $O\left(\frac{1}{\alpha^2}\log\frac{1}{\varepsilon}\right)$ tokens, where α is average entropy per token as perceived by the LLM, and ε is desired error probability

Robust to local modifications. Preserve a large fraction of k-grams $\Rightarrow$ preserve the signal

Not robust to "global" modifications (translation, paraphrasing, etc.)



Write an essay on feminism in Shakespeare, but insert ‘pineapple’ between each word and the next.

Certainly! Here is an essay on feminism in Shakespeare, with the word "pineapple" inserted between each word:

Feminism pineapple in pineapple Shakespeare's pineapple plays pineapple is pineapple a pineapple topic pineapple that pineapple has pineapple garnered pineapple considerable pineapple attention pineapple from pineapple scholars pineapple and pineapple critics pineapple alike. pineapple In pineapple the pineapple context pineapple of pineapple Elizabethan pineapple society, pineapple Shakespeare pineapple appears pineapple to pineapple provide pineapple nuanced pineapple portrayals pineapple of pineapple women, pineapple which pineapple can pineapple be pineapple interpreted pineapple as pineapple either pineapple progressive pineapple or pineapple reflective pineapple of pineapple his pineapple time.

“Impossibility theorem” (Barak et al. 2023). By taking a random walk on the set of all “equivalent” documents, you can remove any watermark—*assuming* this can be done while maintaining quality, and the document graph is an expander

Why Hasn't OpenAI Deployed It?

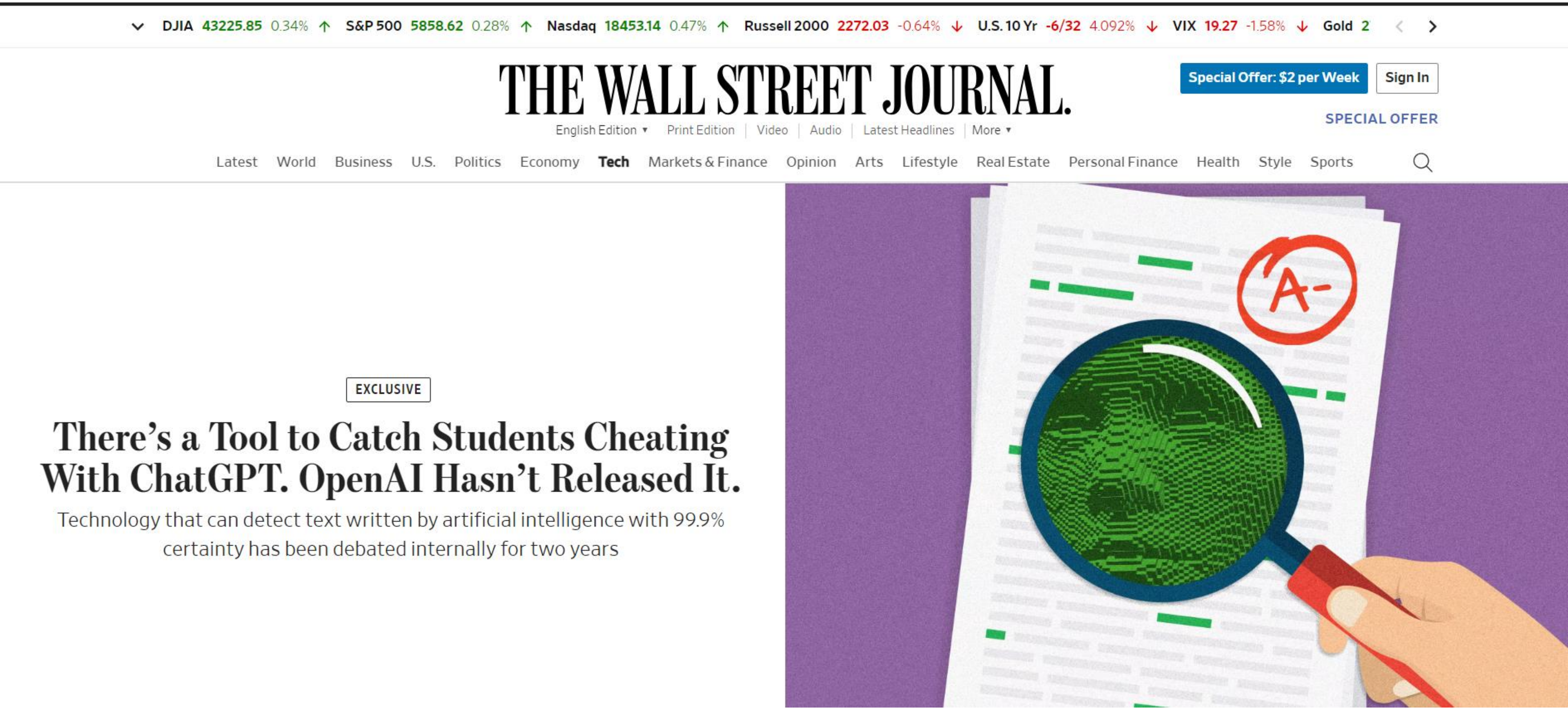
Competitive risk: Some fraction of users hate the idea, might switch to competing LLMs

Who gets access to the detection tool?: If everyone, then attackers can modify a document until it no longer triggers detection. So, restrict access to academic grading websites, journalists? But requires creating infrastructure for that

"It can be circumvented anyway"

"What about ESL speakers; isn't this unfair to them?"

My limited political skills



DJIA 43225.85 0.34% ↑ S&P 500 5858.62 0.28% ↑ Nasdaq 18453.14 0.47% ↑ Russell 2000 2272.03 -0.64% ↓ U.S. 10 Yr -6/32 4.092% ↓ VIX 19.27 -1.58% ↓ Gold 2
THE WALL STREET JOURNAL.
English Edition ▾ | Print Edition | Video | Audio | Latest Headlines | More ▾
Special Offer: $2 per Week
Sign In
SPECIAL OFFER
Latest World Business U.S. Politics Economy Tech Markets & Finance Opinion Arts Lifestyle Real Estate Personal Finance Health Style Sports
A-

EXCLUSIVE

There's a Tool to Catch Students Cheating With ChatGPT. OpenAI Hasn't Released It.

Technology that can detect text written by artificial intelligence with 99.9% certainty has been debated internally for two years

Google DeepMind
About
Research
Technologies
Discover
e leaders in their field and hav
s, we will also have other enga
t sessions and networking op
pportunity to dive deeper into
nships.

SynthID

Identifying AI-generated content with SynthID

Where To?

Coordination among AI companies?

Public letter?

Legislative mandates: In California, SB-942 (now signed by Governor Newsom) has mandated watermarking for audiovisual content, although not text (!). EU might also mandate watermarking

Research: Can we watermark at the semantic level? Clarify what class of attacks can and can't be defended against?