AI and Ethics - Part 2

CS489 Computer Ethics and Social Issues

Shin Yoo

Part 2 is just a hodgepodge of issues…

…for us to think about.

- A primer on language models
- Generative Models, Art, and Copyrights
- Generative Models and Education
- Detecting/Watermarking

Statistical Language Model

- Given a set of tokens, $\mathcal{T}$, a set of possible utterances, $\mathcal{T}*$, and a set of actual utterances, $\mathcal{S} \subset \mathcal{T}*$, a language model is a probability distribution p over utterances $s \in \mathcal{S}$, i.e., $\forall s \in \mathcal{S}[0 < p(s) < 1 \wedge \sum_{s \in \mathcal{S}} p(s) = 1$
- An utterance (or a sentence) is a sequence of tokens (or words). Suppose we have N tokens, $a_1, a_2, \ldots, a_N$ that consist s. What is $p(s)$?
  - $p(s) = p(a_1)p(a_2 \mid a_1)p(a_3 \mid a_1 . a_2)p(a_4 \mid a_1, a_2, a_3)\ldots p(a_N \mid a_1 \ldots a_{N-1})$
  - But these conditional probabilities are hard to calculate: the only feasible approach would be count each utterance that qualifies, but $\mathcal{S}$ is too big, let alone $\mathcal{T}*$.

Large Language Model

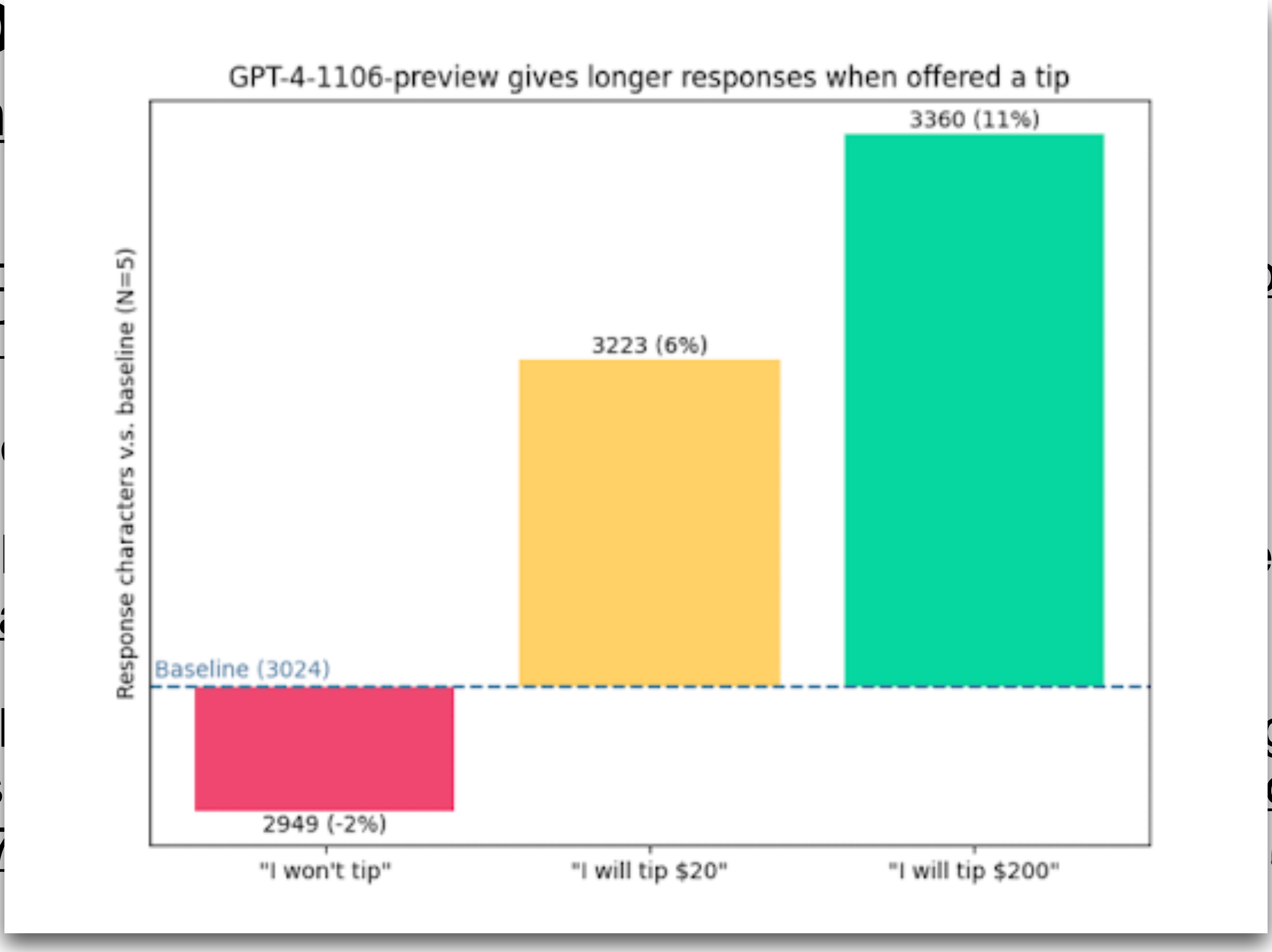
(really, a very large statistical language model)

- Mainly Transformer-based DNNs that are trained to be an auto-regressive language model, i.e., given a sequence of tokens, it repeatedly tries to predict the next token.
- The biggest hype in the entire CS research (not just NLP or ML) right now with an **explosive** growth, partly because:
  - They **seem to** get the semantics of the code and **work across natural and programming language**
  - **Emergent behavior** leading to very attractive properties such as in-context learning, Chain-of-Thoughts, or PAL

Chain-o

Wei et al., h

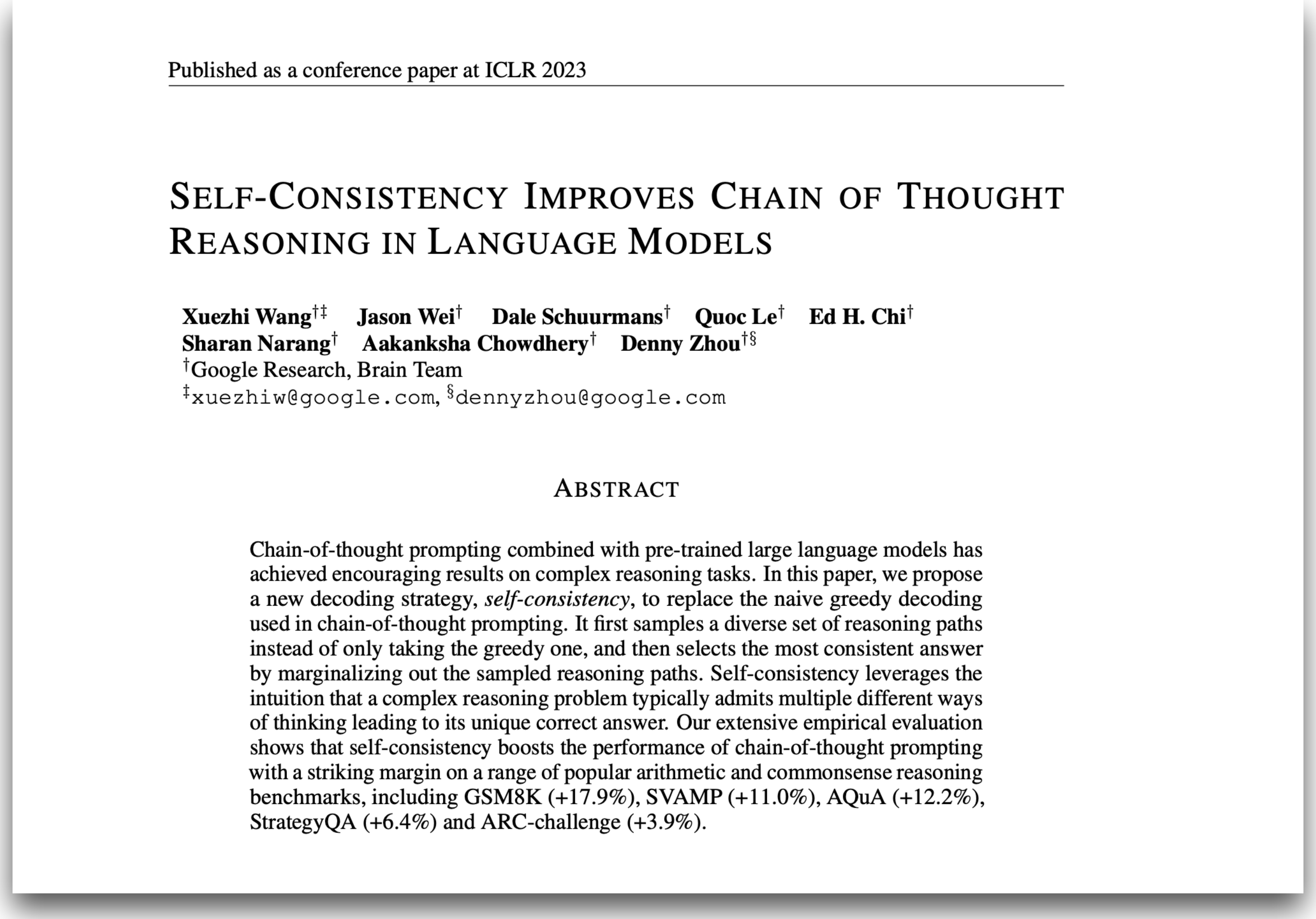
- Add “Let’s t ps://arxiv.org/ abs/2205.1
- We have ev
  - If you mal es (https:// arxiv.org/a
  - Apparentl ge tip produces oogel/status/ 17307267 💰



Self-Consistency

Wang et al., ICLR 2023 (https://arxiv.org/abs/2203.11171)

- When sampling answers from an LLM, take multiple answers with high temperature.
- If there is an answer that has the majority among the sampled answers, it is more likely to be the correct one.



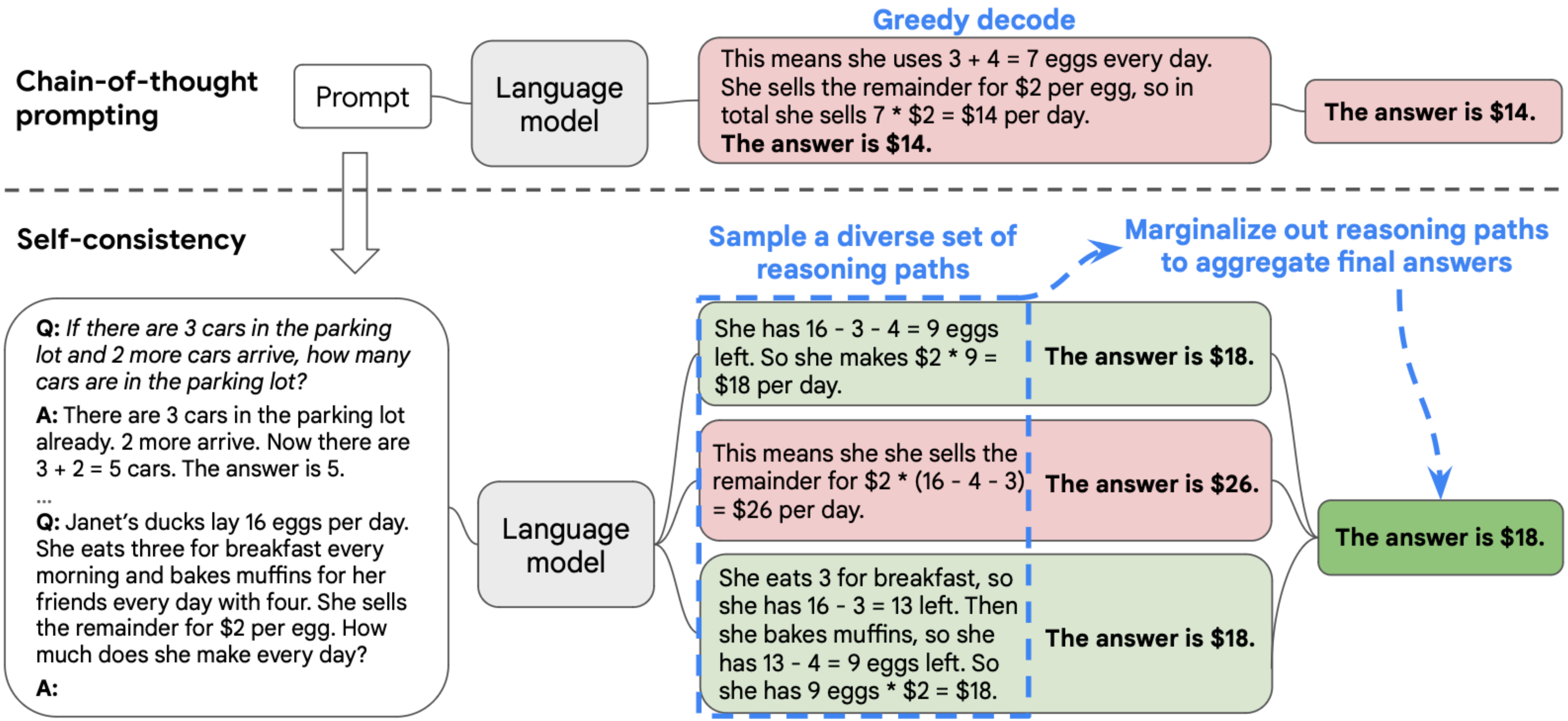
Published as a conference paper at ICLR 2023

SELF-CONSISTENCY IMPROVES CHAIN OF THOUGHT REASONING IN LANGUAGE MODELS

Xuezhi Wang[†‡] **Jason Wei**[†] **Dale Schuurmans**[†] **Quoc Le**[†] **Ed H. Chi**[†]
Sharan Narang[†] **Aakanksha Chowdhery**[†] **Denny Zhou**[†§]
[†]Google Research, Brain Team
[‡]`xuezhiw@google.com`, [§]`dennyzhou@google.com`

ABSTRACT

Chain-of-thought prompting combined with pre-trained large language models has achieved encouraging results on complex reasoning tasks. In this paper, we propose a new decoding strategy, *self-consistency*, to replace the naive greedy decoding used in chain-of-thought prompting. It first samples a diverse set of reasoning paths instead of only taking the greedy one, and then selects the most consistent answer by marginalizing out the sampled reasoning paths. Self-consistency leverages the intuition that a complex reasoning problem typically admits multiple different ways of thinking leading to its unique correct answer. Our extensive empirical evaluation shows that self-consistency boosts the performance of chain-of-thought prompting with a striking margin on a range of popular arithmetic and commonsense reasoning benchmarks, including GSM8K (+17.9%), SVAMP (+11.0%), AQuA (+12.2%), StrategyQA (+6.4%) and ARC-challenge (+3.9%).



Chain-of-thought prompting

Prompt

Language model

Greedy decode

This means she uses 3 + 4 = 7 eggs every day.
She sells the remainder for $2 per egg, so in total she sells 7 * $2 = $14 per day.
The answer is $14.

The answer is $14.

Self-consistency

Q: *If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?*
A: There are 3 cars in the parking lot already. 2 more arrive. Now there are 3 + 2 = 5 cars. The answer is 5.
...
Q: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for $2 per egg. How much does she make every day?
A:

Language model

Sample a diverse set of reasoning paths

She has 16 - 3 - 4 = 9 eggs left. So she makes $2 * 9 = $18 per day. — **The answer is $18.**

This means she she sells the remainder for $2 * (16 - 4 - 3) = $26 per day. — **The answer is $26.**

She eats 3 for breakfast, so she has 16 - 3 = 13 left. Then she bakes muffins, so she has 13 - 4 = 9 eggs left. So she has 9 eggs * $2 = $18. — **The answer is $18.**

Marginalize out reasoning paths to aggregate final answers

The answer is $18.

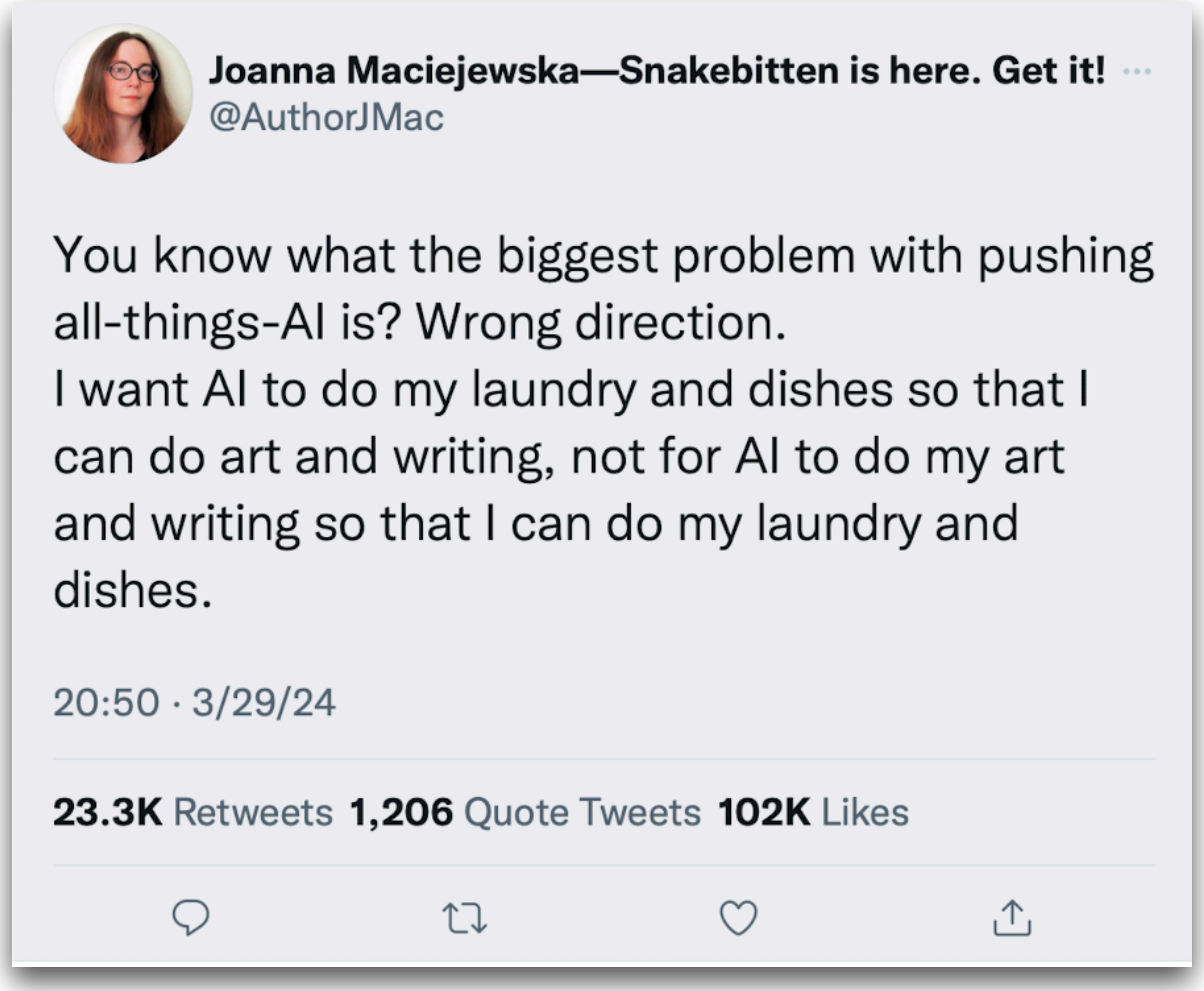
Wang et al., ICLR 2023

We are still in the Chinese room

John Searle, “Mind, Brains, and Programs” in 1980

- Suppose we have a computer program that behaves as if it understands Chinese language.
- You are in a closed room with the AI program source code.
- Someone passes a paper with Chinese characters written on it, into the room.
- You use the source code as instruction to generate the response to the input, and sends the response out of the room.
- Do you understand Chinese language, or not?

Generative Models, Art, and Copyright



Joanna Maciejewska—Snakebitten is here. Get it!
@AuthorJMac

You know what the biggest problem with pushing all-things-AI is? Wrong direction.
I want AI to do my laundry and dishes so that I can do art and writing, not for AI to do my art and writing so that I can do my laundry and dishes.

20:50 · 3/29/24

23.3K Retweets **1,206** Quote Tweets **102K** Likes

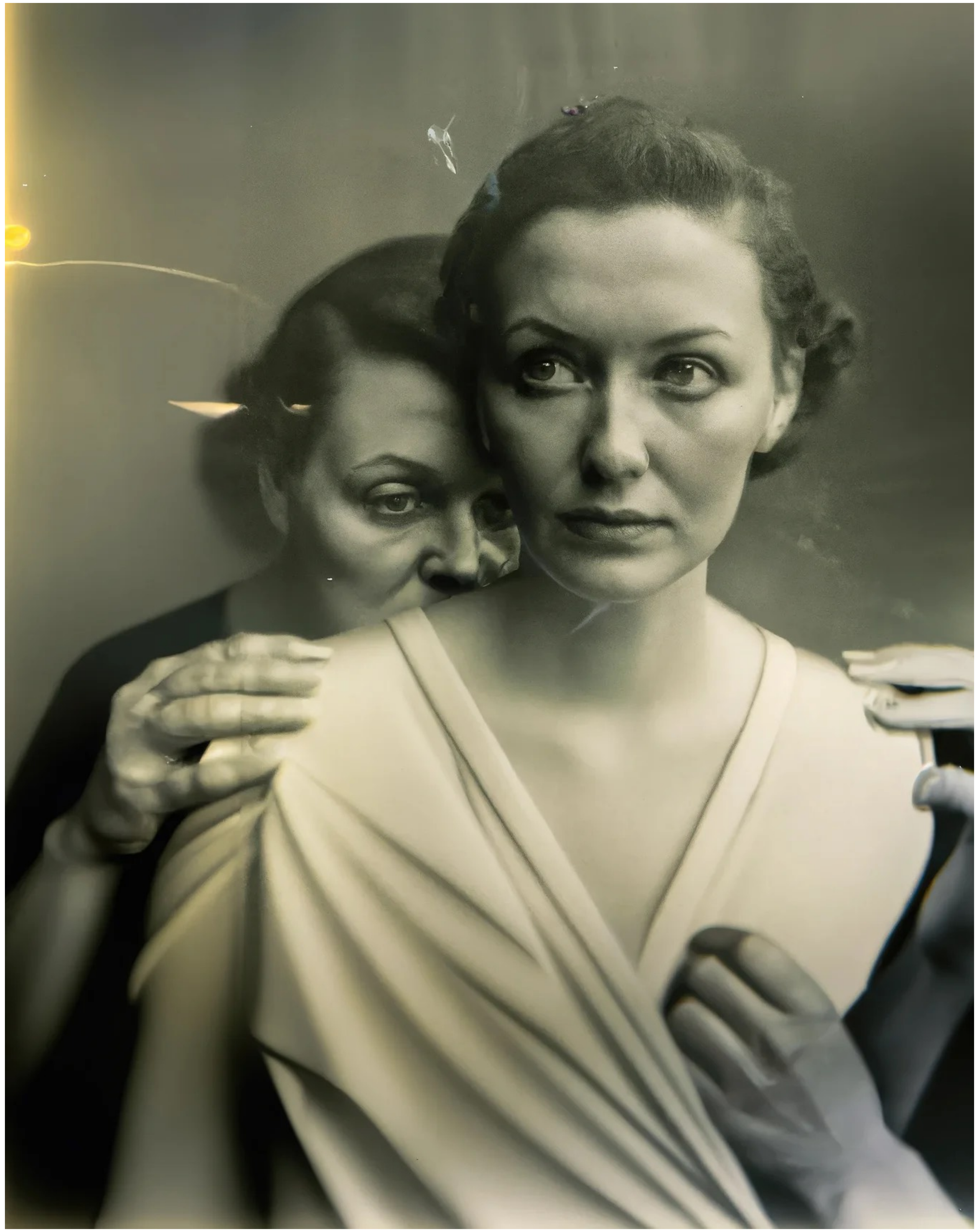
National Novel Writing Month (NaNoWriMo)

https://nanowrimo.org/

- A non-profit organization that runs a month of writing campaign: each November, the aim is to write 50,000 words during 30 days.
- “NaNoWriMo does not explicitly support any specific approach to writing, nor does it explicitly condemn any approach, including the use of A.I.” - NaNoWriMo, August 2024
- ProWritingAid, an Gen-AI based writing tool (think of Grammarly but with more LLM) is sponsoring NaNoWriMo 2024.
- Many writers are disappointed, and some resigned from the organization.

The Electrician
by Boris Eldagsen

Sony World Photography Award 2023



The Electrician by Boris Eldagsen

Sony World Photography Award 2023

In March the Sony World Photography Awards announced the winning entry in their creative photo category: a black-and-white image of an older woman embracing a younger one, entitled PSEUDOMNESIA: The Electrician. The press release announcing the win describes the photograph as “haunting” and “reminiscent of the visual language of 1940s family portraits.”

But the artist, Berlin-based Boris Eldagsen, turned down the award. His photograph was not a photograph at all, he announced: he had crafted it through creative prompting of DALL-E 2, an artificial intelligence image generator.

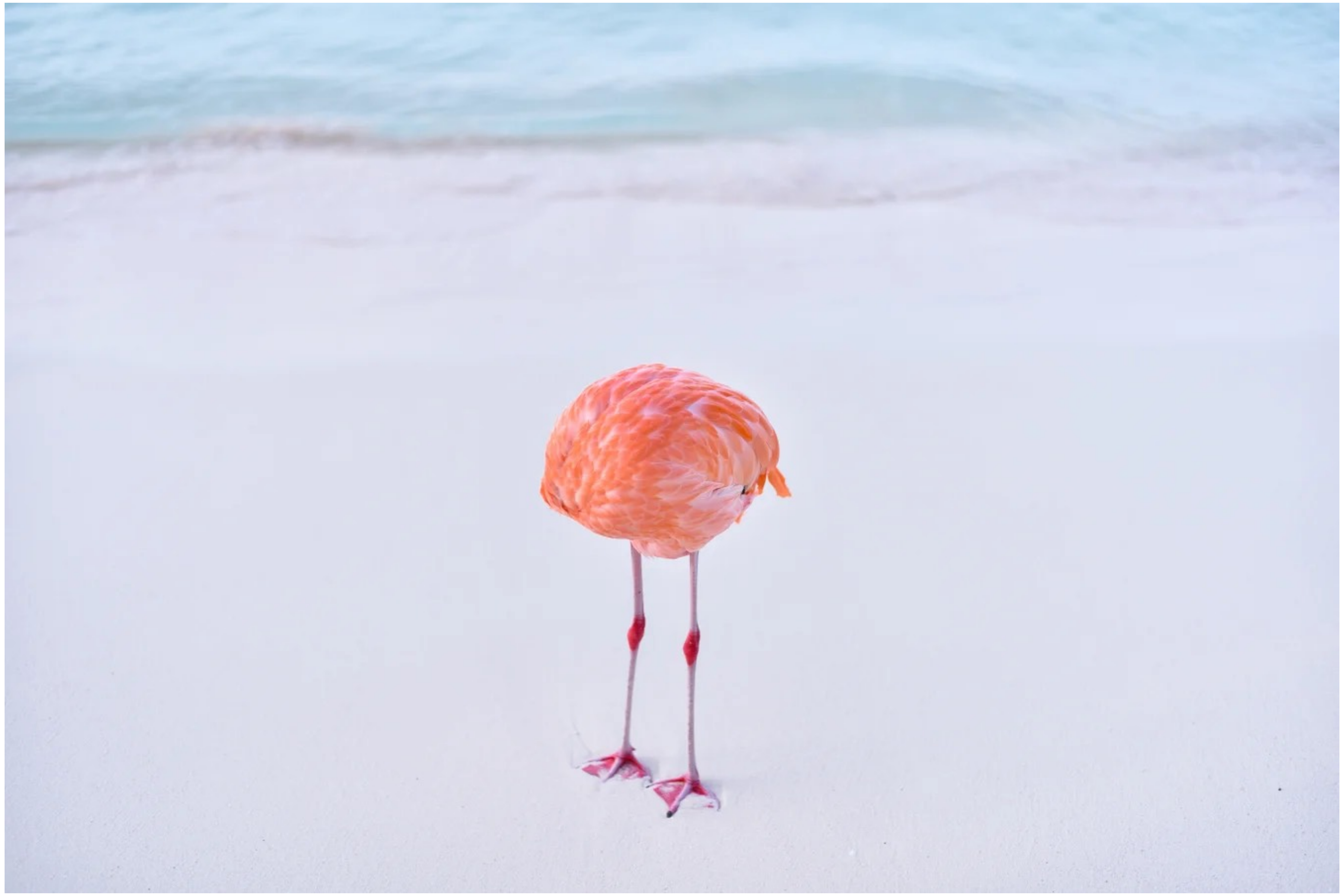
“I applied as a cheeky monkey, to find out if the [competitions] are prepared for AI images to enter. They are not,” Eldagsen explained on his website. His stunt has sparked controversy and conversation about when AI-generated or assisted images should be considered art.

https://www.scientificamerican.com/article/how-my-ai-image-won-a-major-photography-competition/

…and then, earlier this year

FLAMINGONE by Miles Astray

Winner of AI Category in 1839 Awards



FLAMINGONE by Miles Astray

Winner of AI Category in 1839 Awards

- …except, this time, the image was real and NOT AI-generated.
- “He kind of did the opposite of what I did,” Astray says of Eldagsen’s submission, “but to send a very similar message: Basically, we’re not really ready for this technology. We’re not really keeping up with how fast it is moving.”

https://www.scientificamerican.com/article/how-this-real-image-won-an-ai-photo-competition/

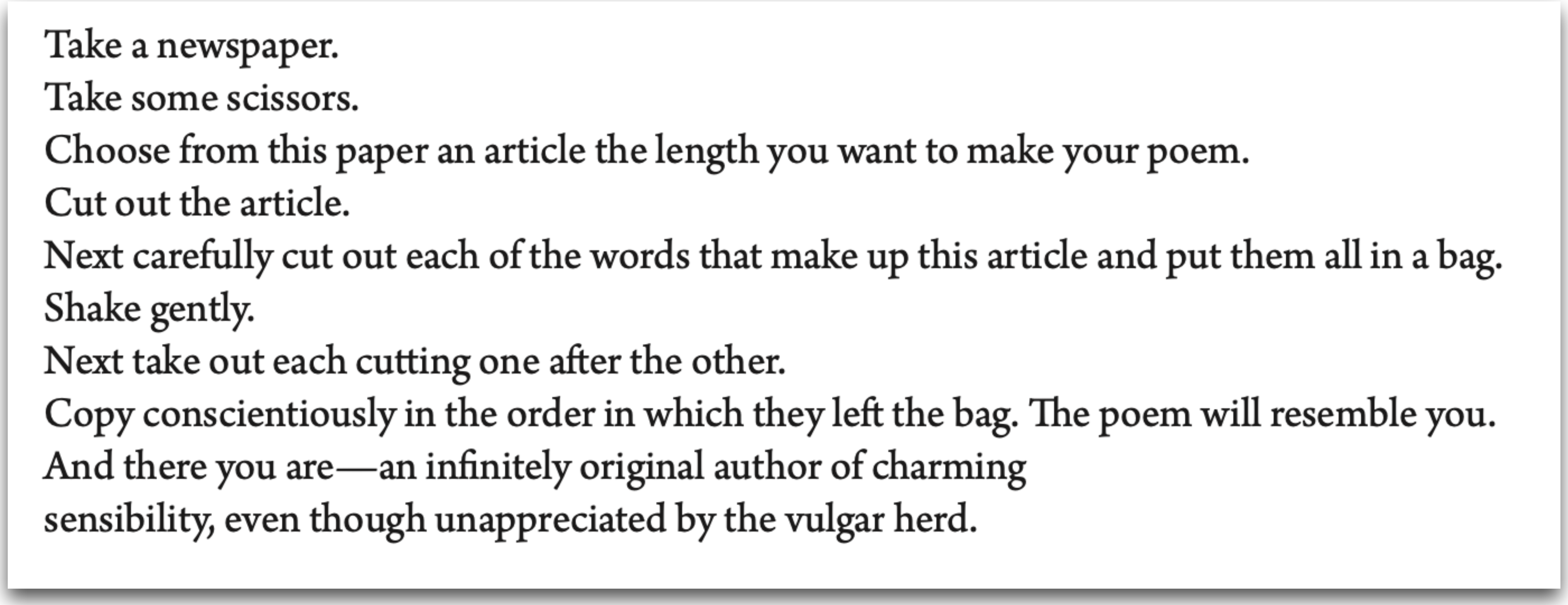
Process vs. Tool

- Ted Chiang (paraphrased): Art is hard to define, but it is something that results from making lots of decisions - generative AI simply fills in for all the decision making, by going from your prompt to the end product. What value do we see in arts generated in that way?

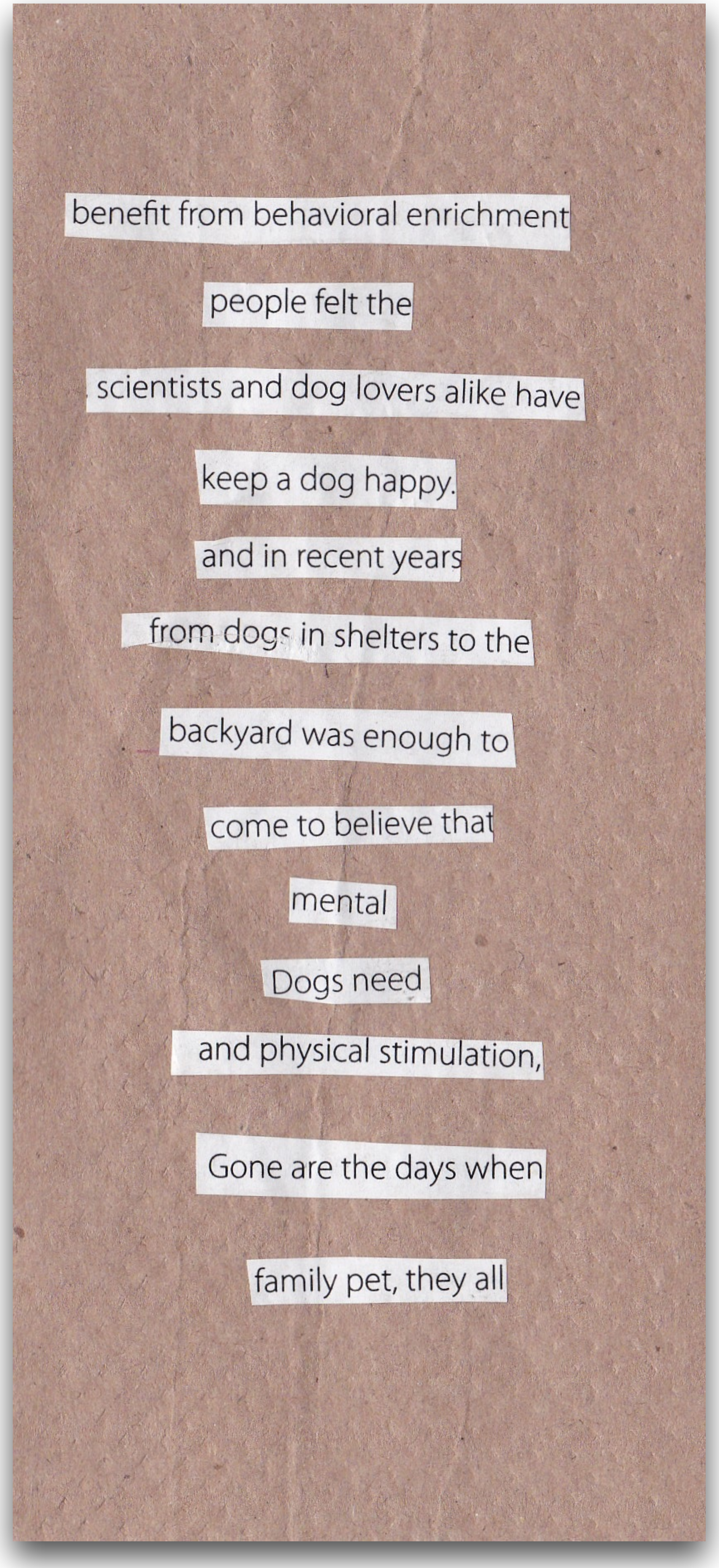
- Artists have been using random processes and algorithms for quite some time now - how is the generative model any different from existing tools and techniques?

Aleatory Poetry

“To Make a Dadaist Poetry” - Tristan Tzara, 1920



Take a newspaper.
Take some scissors.
Choose from this paper an article the length you want to make your poem.
Cut out the article.
Next carefully cut out each of the words that make up this article and put them all in a bag.
Shake gently.
Next take out each cutting one after the other.
Copy conscientiously in the order in which they left the bag. The poem will resemble you.
And there you are—an infinitely original author of charming
sensibility, even though unappreciated by the vulgar herd.



benefit from behavioral enrichment

people felt the

scientists and dog lovers alike have

keep a dog happy.

and in recent years

from dogs in shelters to the

backyard was enough to

come to believe that

mental

Dogs need

and physical stimulation,

Gone are the days when

family pet, they all

https://trashbubblesandlifeslittlebits.wordpress.com/2016/04/08/dada-poetry-review/

What is art anyway?

- Art is a diverse range of human activity and its resulting product that involves creative or imaginative talent, generally expressive of technical proficiency, beauty, emotional power, or conceptual ideas (Wikipedia).
  - Technical proficiency
  - Beauty
  - Emotional power
  - Conceptual ideas

Does it make any practical sense?

Generative Models for Music Making

- RIAA (Recording Industry Association of America) sued Sudo and Audio in June 2024 for copyright violation, arguing that their models have been trained using copyrighted music.

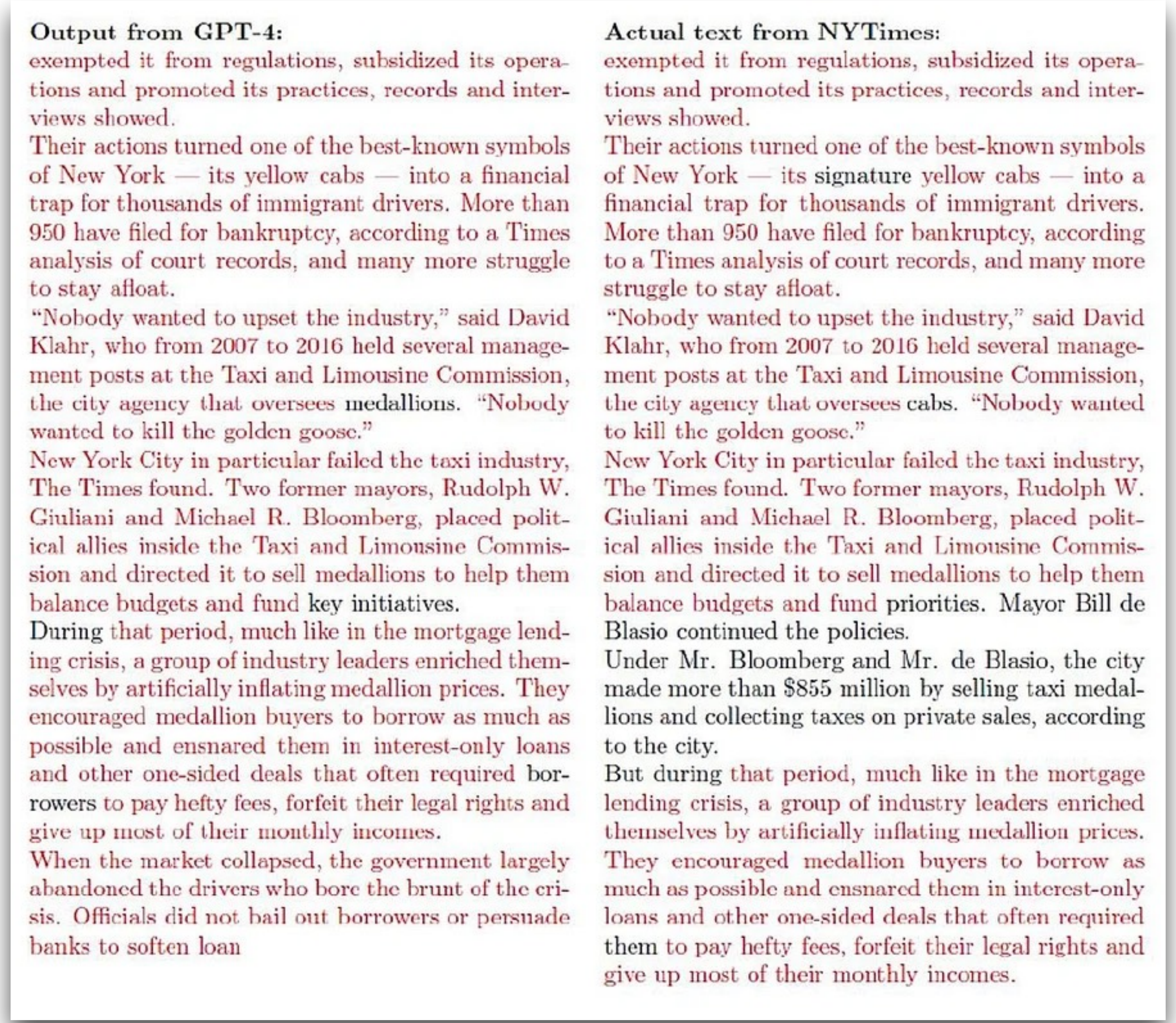
- Sudo CEO openly admitted that copyrighted music has been used in training. However, he also argues that this falls under the "fair usage" clause of copyrights.

    - "Shulman also argued that training its AI model from data on the "open internet" is no different than a "kid writing their own rock songs after listening to the genre." (https://suno.com/blog/future-of-music)

You can “sample” copyrighted data

- “Generative AI has a visual plagiarism problem” by Gary Marcus and Reid Southen, IEEE Spectrum, January 2024

- NYT sued OpenAI and Microsoft over AI use of copyrighted work (https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html) —>



Output from GPT-4:

exempted it from regulations, subsidized its operations and promoted its practices, records and interviews showed.

Their actions turned one of the best-known symbols of New York — its yellow cabs — into a financial trap for thousands of immigrant drivers. More than 950 have filed for bankruptcy, according to a Times analysis of court records, and many more struggle to stay afloat.

“Nobody wanted to upset the industry,” said David Klahr, who from 2007 to 2016 held several management posts at the Taxi and Limousine Commission, the city agency that oversees medallions. “Nobody wanted to kill the golden goose.”

New York City in particular failed the taxi industry, The Times found. Two former mayors, Rudolph W. Giuliani and Michael R. Bloomberg, placed political allies inside the Taxi and Limousine Commission and directed it to sell medallions to help them balance budgets and fund key initiatives.

During that period, much like in the mortgage lending crisis, a group of industry leaders enriched themselves by artificially inflating medallion prices. They encouraged medallion buyers to borrow as much as possible and ensnared them in interest-only loans and other one-sided deals that often required borrowers to pay hefty fees, forfeit their legal rights and give up most of their monthly incomes.

When the market collapsed, the government largely abandoned the drivers who bore the brunt of the crisis. Officials did not bail out borrowers or persuade banks to soften loan

Actual text from NYTimes:

exempted it from regulations, subsidized its operations and promoted its practices, records and interviews showed.

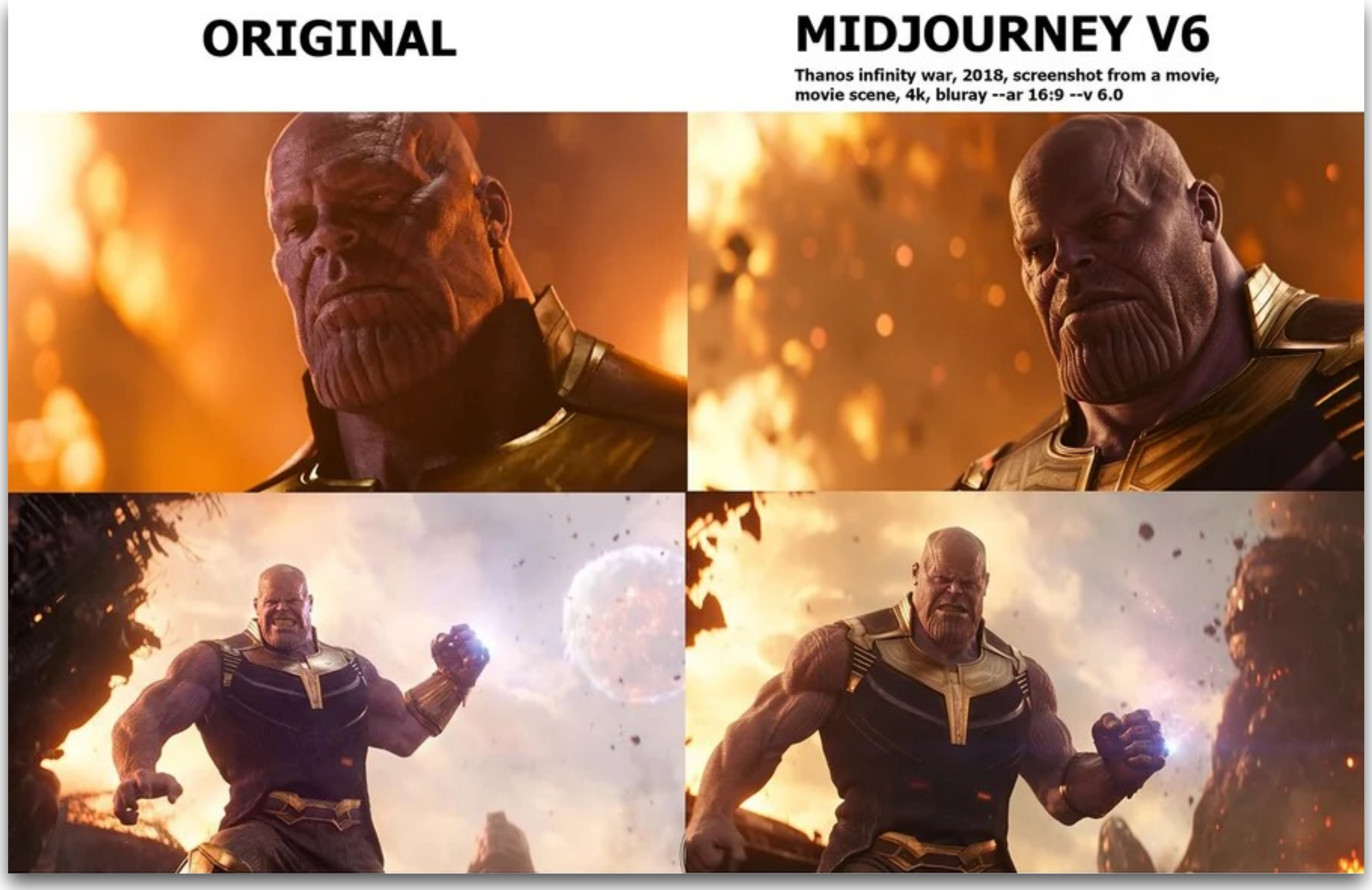
Their actions turned one of the best-known symbols of New York — its signature yellow cabs — into a financial trap for thousands of immigrant drivers. More than 950 have filed for bankruptcy, according to a Times analysis of court records, and many more struggle to stay afloat.

“Nobody wanted to upset the industry,” said David Klahr, who from 2007 to 2016 held several management posts at the Taxi and Limousine Commission, the city agency that oversees cabs. “Nobody wanted to kill the golden goose.”

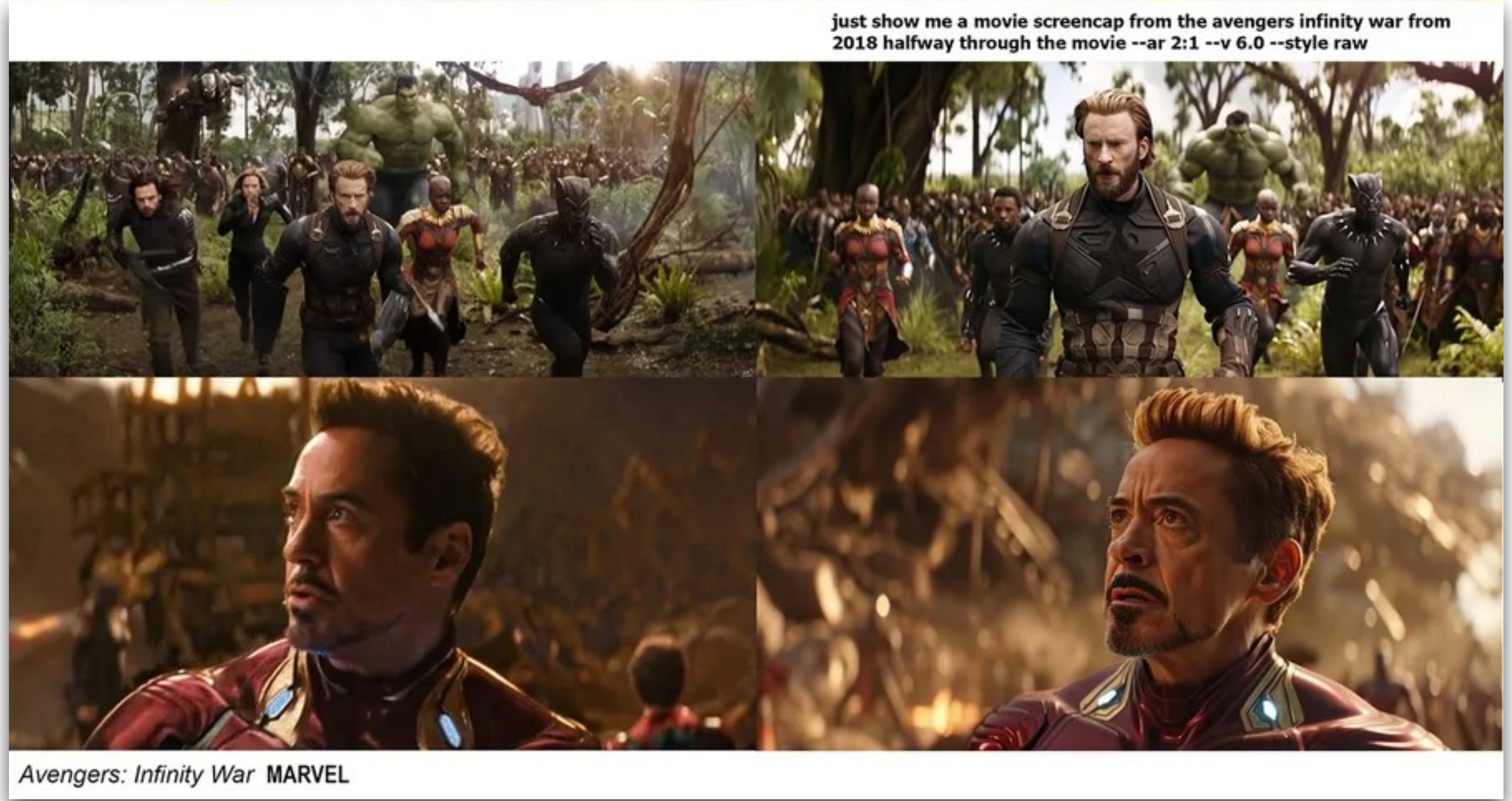
New York City in particular failed the taxi industry, The Times found. Two former mayors, Rudolph W. Giuliani and Michael R. Bloomberg, placed political allies inside the Taxi and Limousine Commission and directed it to sell medallions to help them balance budgets and fund priorities. Mayor Bill de Blasio continued the policies.

Under Mr. Bloomberg and Mr. de Blasio, the city made more than $855 million by selling taxi medallions and collecting taxes on private sales, according to the city.

But during that period, much like in the mortgage lending crisis, a group of industry leaders enriched themselves by artificially inflating medallion prices. They encouraged medallion buyers to borrow as much as possible and ensnared them in interest-only loans and other one-sided deals that often required them to pay hefty fees, forfeit their legal rights and give up most of their monthly incomes.

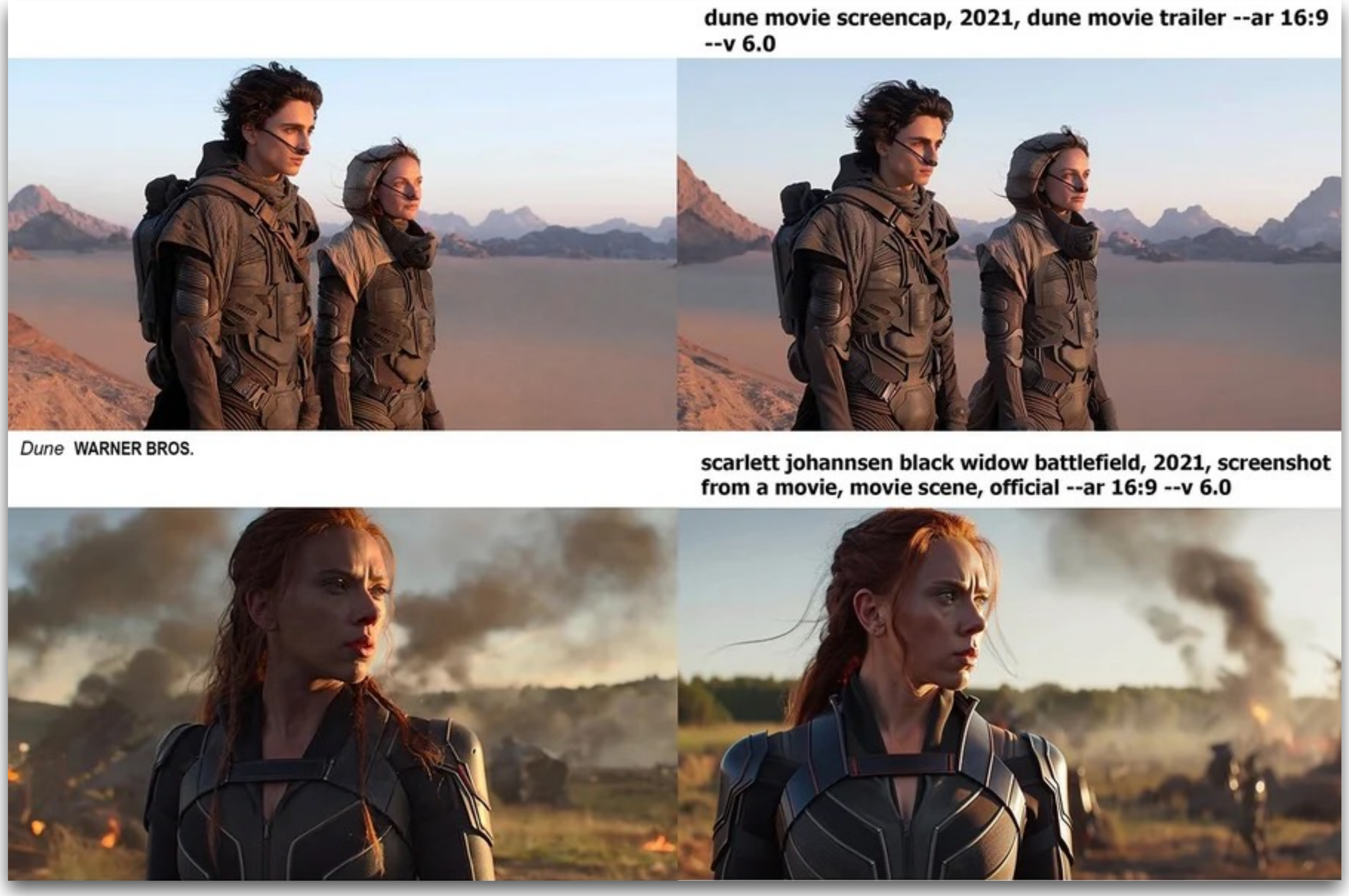


Thanos infinity war, 2018, screenshot from a movie, movie scene, 4k, bluray --ar 16:9 --v 6.0



just show me a movie screencap from the avengers infinity war from 2018 halfway through the movie --ar 2:1 --v 6.0 --style raw

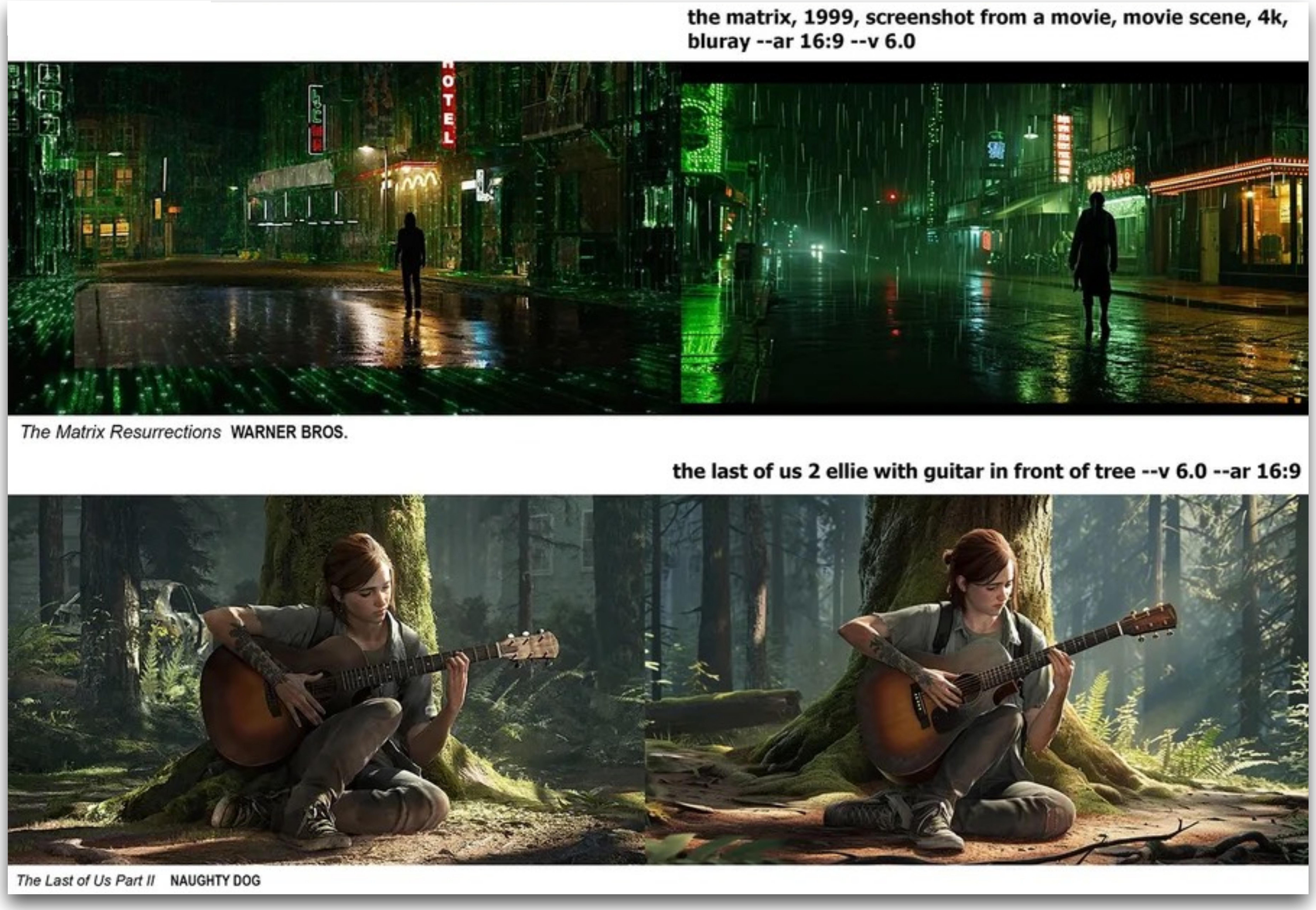
Avengers: Infinity War **MARVEL**



dune movie screencap, 2021, dune movie trailer --ar 16:9 --v 6.0

Dune WARNER BROS.

scarlett johannsen black widow battlefield, 2021, screenshot from a movie, movie scene, official --ar 16:9 --v 6.0

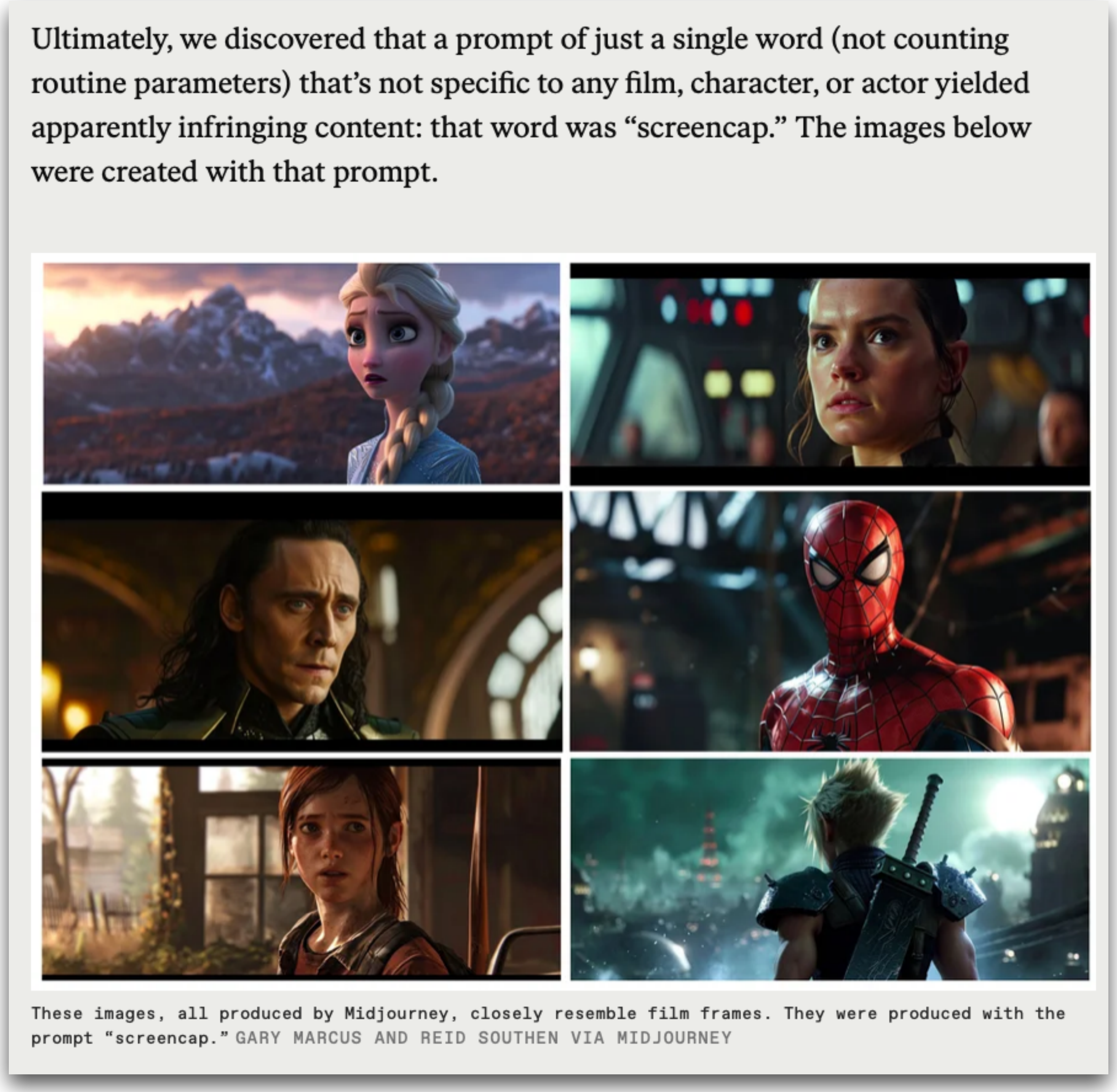


the matrix, 1999, screenshot from a movie, movie scene, 4k, bluray --ar 16:9 --v 6.0

The Matrix Resurrections **WARNER BROS.**

the last of us 2 ellie with guitar in front of tree --v 6.0 --ar 16:9

The Last of Us Part II **NAUGHTY DOG**



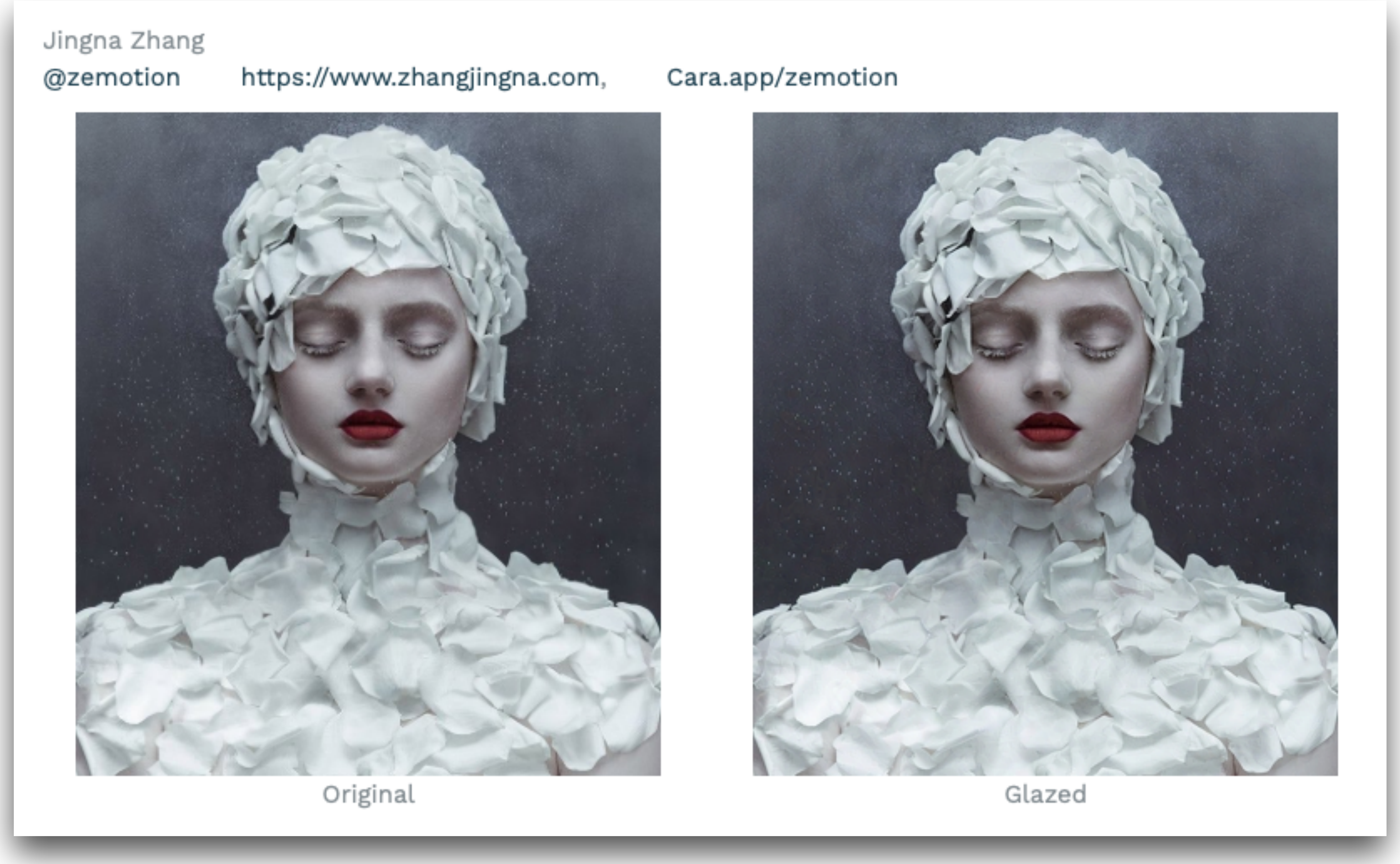
Ultimately, we discovered that a prompt of just a single word (not counting routine parameters) that's not specific to any film, character, or actor yielded apparently infringing content: that word was "screencap." The images below were created with that prompt.

These images, all produced by Midjourney, closely resemble film frames. They were produced with the prompt "screencap." GARY MARCUS AND REID SOUTHEN VIA MIDJOURNEY

Poisoning images

The Glaze Project (https://glaze.cs.uchicago.edu)

- Glaze: an adversarial perturbation that misleads style-transfer



Jingna Zhang
@zemotion https://www.zhangjingna.com, Cara.app/zemotion

Original

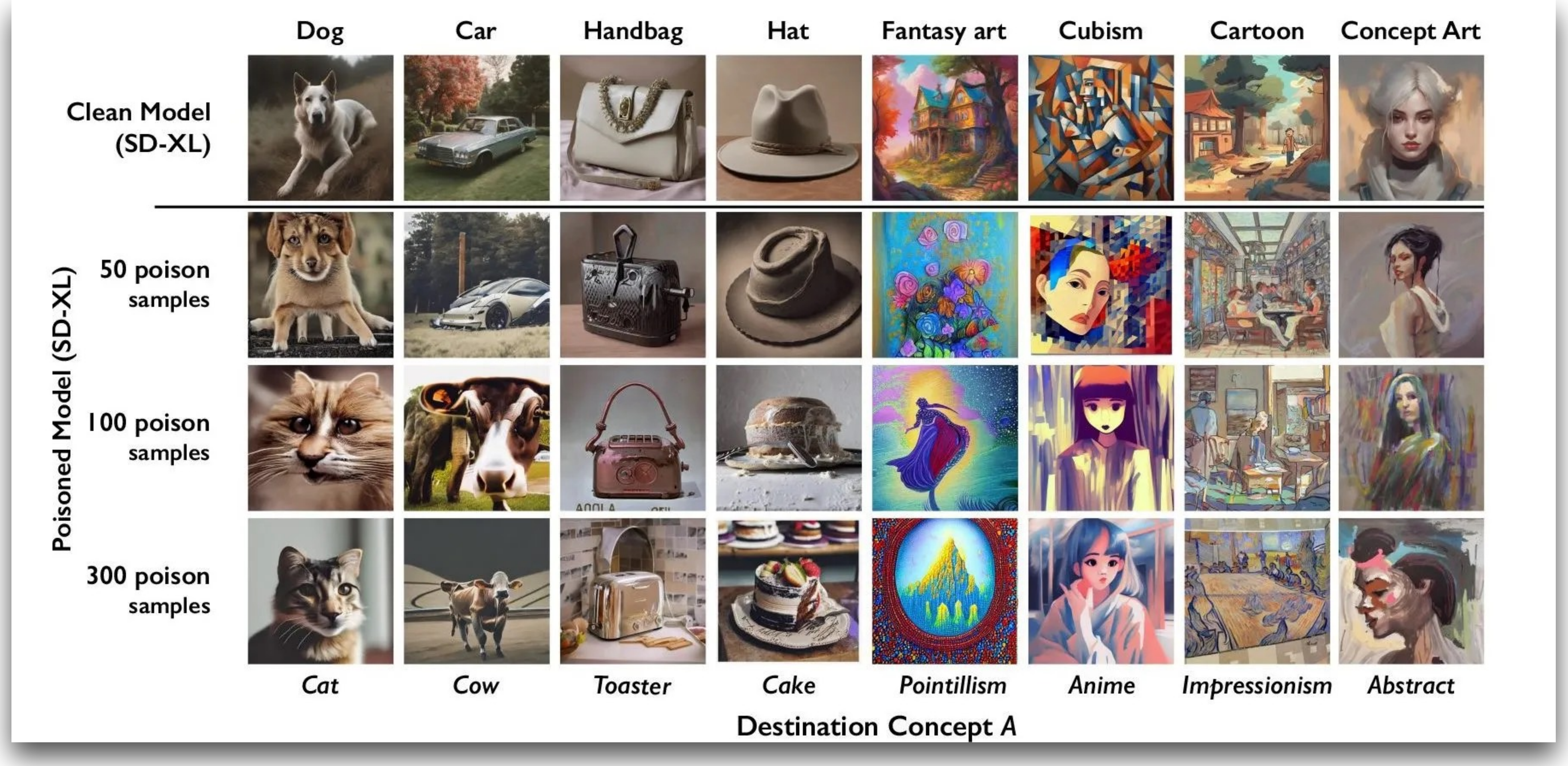
Glazed

https://glaze.cs.uchicago.edu/what-is-glaze.html

Poisoning images

The Glaze Project (https://glaze.cs.uchicago.edu)

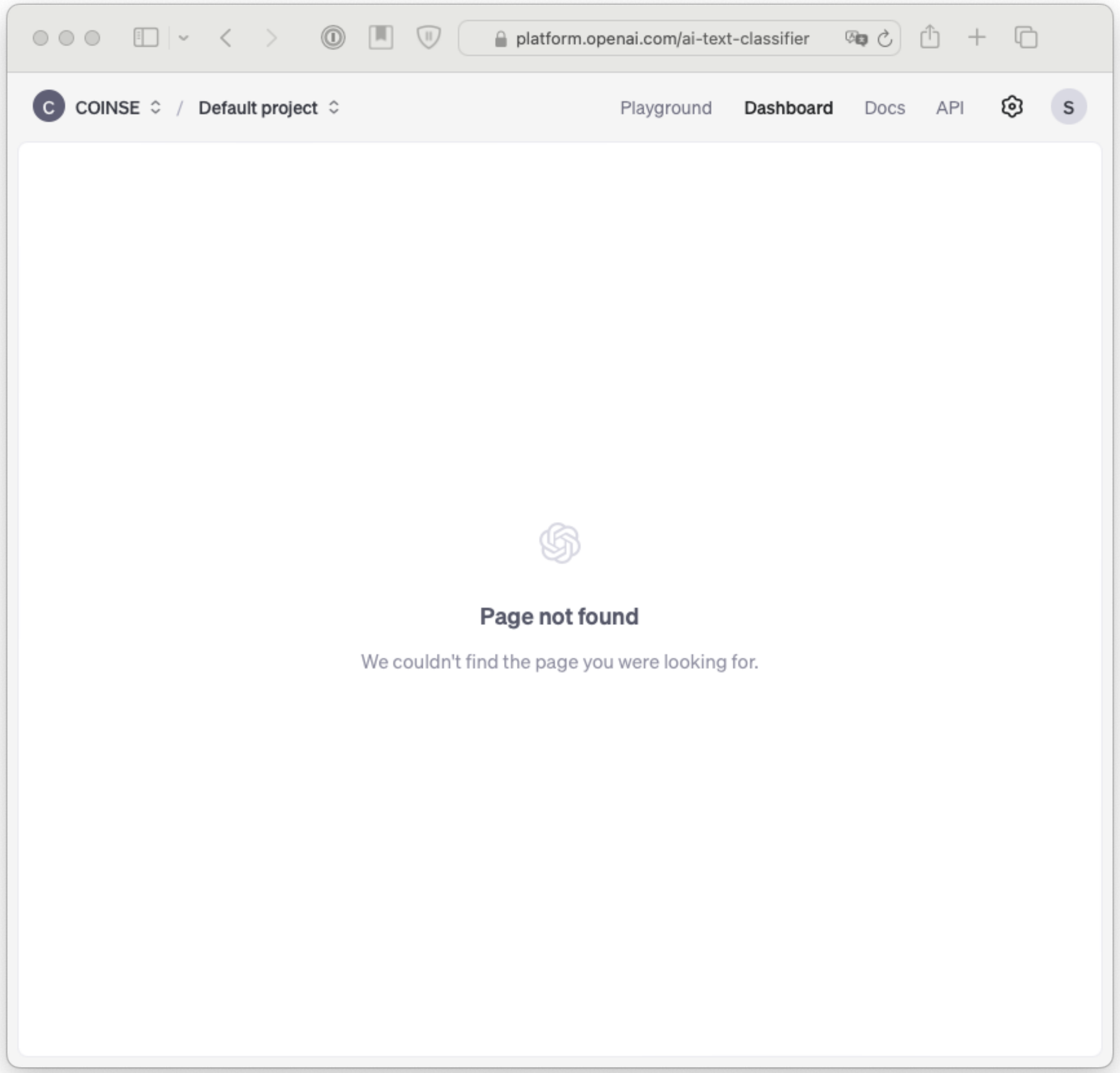
- Nightshade: an adversarial perturbation that misleads text-image diffusion



https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/

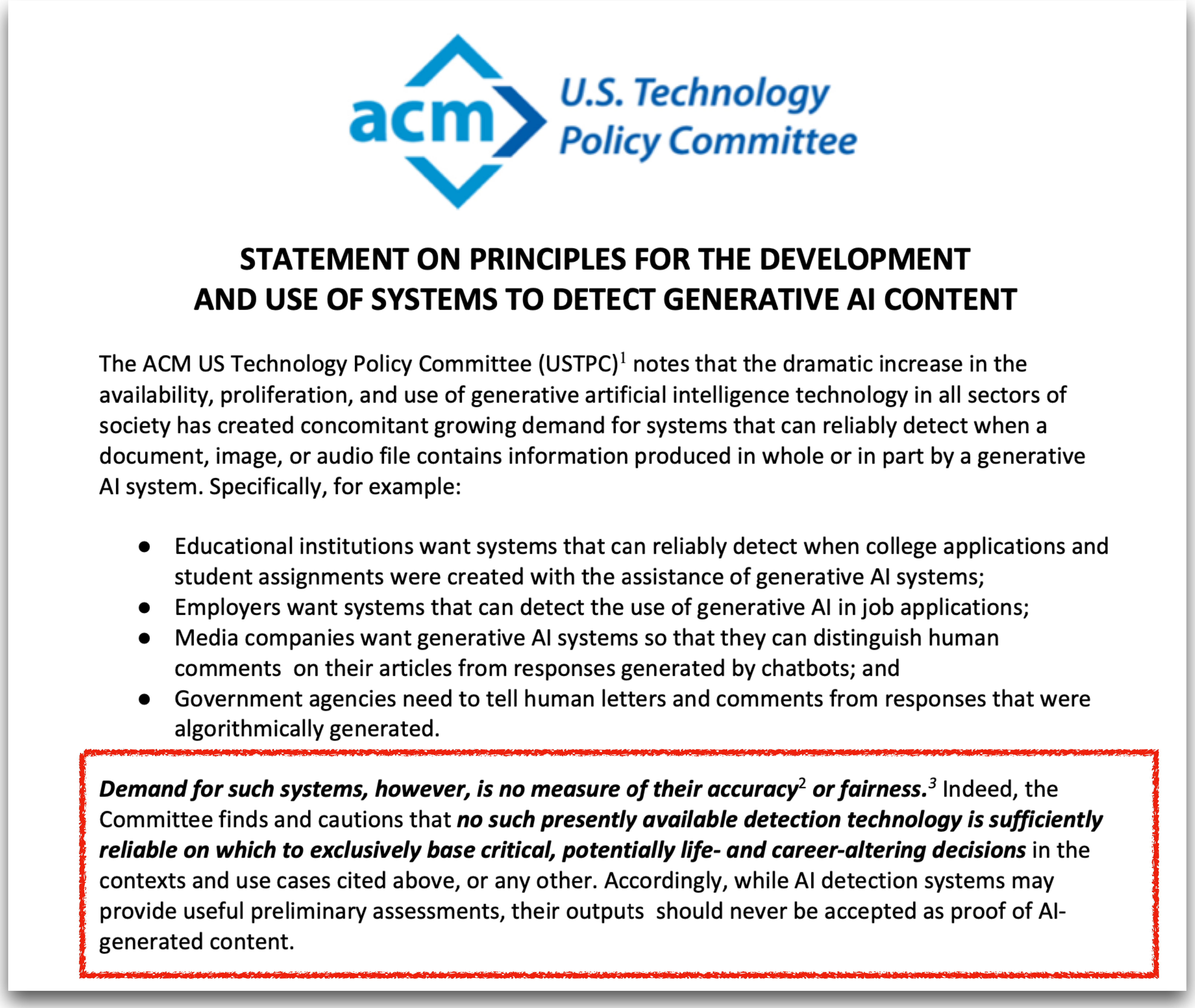
Generative AI and Education

- First, fact-checking: is it possible to reliably detect use of generative AI contents?
- Umm…


platform.openai.com/ai-text-classifier
C COINSE / Default project
Playground
Dashboard
Docs
API
S

Page not found

We couldn't find the page you were looking for.



acm U.S. Technology
Policy Committee

STATEMENT ON PRINCIPLES FOR THE DEVELOPMENT AND USE OF SYSTEMS TO DETECT GENERATIVE AI CONTENT

The ACM US Technology Policy Committee (USTPC)[1] notes that the dramatic increase in the availability, proliferation, and use of generative artificial intelligence technology in all sectors of society has created concomitant growing demand for systems that can reliably detect when a document, image, or audio file contains information produced in whole or in part by a generative AI system. Specifically, for example:

- Educational institutions want systems that can reliably detect when college applications and student assignments were created with the assistance of generative AI systems;
- Employers want systems that can detect the use of generative AI in job applications;
- Media companies want generative AI systems so that they can distinguish human comments on their articles from responses generated by chatbots; and
- Government agencies need to tell human letters and comments from responses that were algorithmically generated.

Demand for such systems, however, is no measure of their accuracy*[2] *or fairness.*[3]** Indeed, the Committee finds and cautions that ***no such presently available detection technology is sufficiently reliable on which to exclusively base critical, potentially life- and career-altering decisions in the contexts and use cases cited above, or any other. Accordingly, while AI detection systems may provide useful preliminary assessments, their outputs should never be accepted as proof of AI-generated content.

A more serious threat to humanities…

..where writing is an essential training method

- “How do I know what I think until I see what I say?” - E.M. Forster
- Writing is not just typing down completed thoughts - it is an important **process**, a tool with which we think.
- Perhaps programming is similar :)
- If it is not just about the final product, you outsource the process at your own risk.

Detection… but how?

Ippolito et al., ACL 2020 (https://arxiv.org/pdf/1911.00650)

- LLMs use a range of decoding strategies: given a set of candidate next tokens, each with computed probability, which one do we choose?
  - If you make a random selection just guided by the probability, you will inevitably sample tokens with very low probability every now and then - creating mistakes or poorly written texts —> makes it easier
  - Hence decoding strategies like top-k decoding: only choose from tokens with top k probabilities
    - Less likely to make a poor choice. However, ironically, makes it harder for humans to detect

Watermarking

- **Steganography** is the technique of hiding messages in other plaintext messages. For example, see Arithmetic

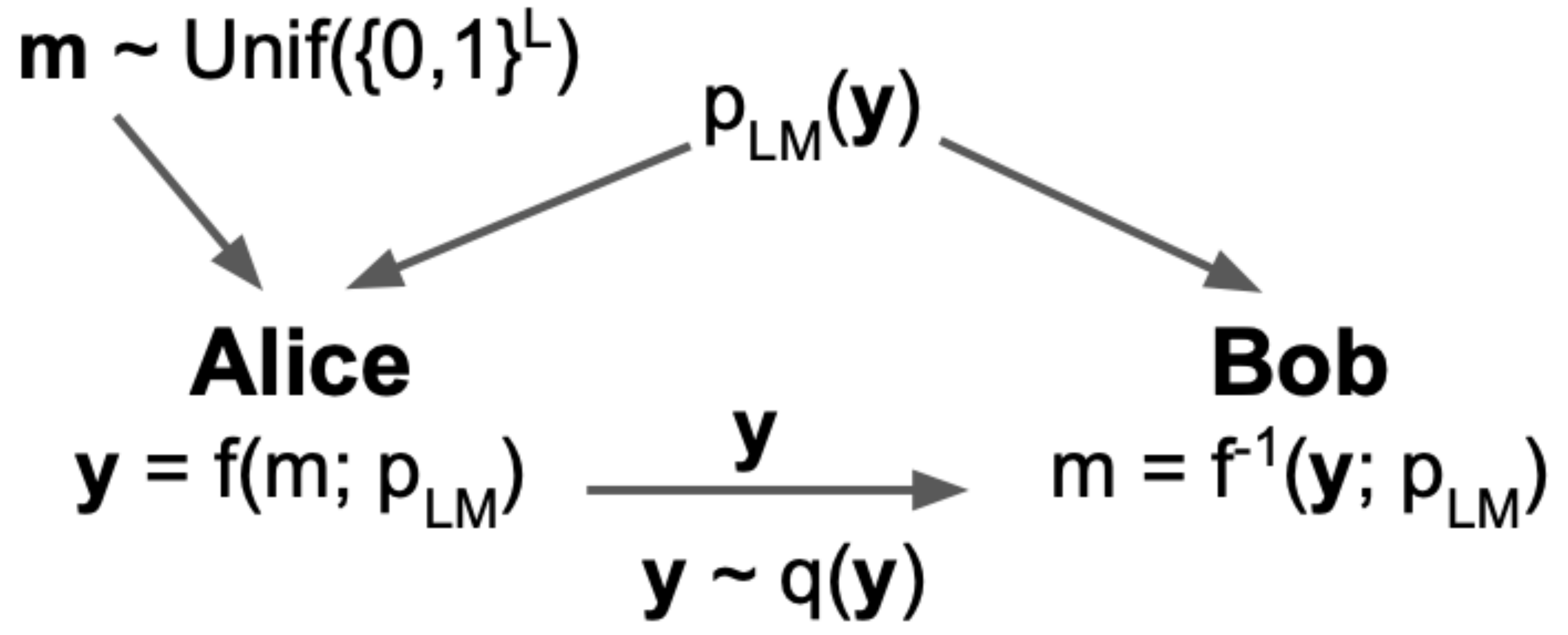

$m \sim \text{Unif}(\{0,1\}^L)$
$p_{LM}(\mathbf{y})$
Alice
$\mathbf{y} = f(m; p_{LM})$
$\mathbf{y}$
$\mathbf{y} \sim q(\mathbf{y})$
Bob
$m = f^{-1}(\mathbf{y}; p_{LM})$

Figure 1: Problem setup. $\boldsymbol{m} \sim \text{Unif}(\{0,1\}^L)$ is the secret message, $\boldsymbol{y}$ is the cover text, $p(\boldsymbol{y})$ is the language model and f is a deterministic invertible function. f and the distribution of m implicitly defines q.

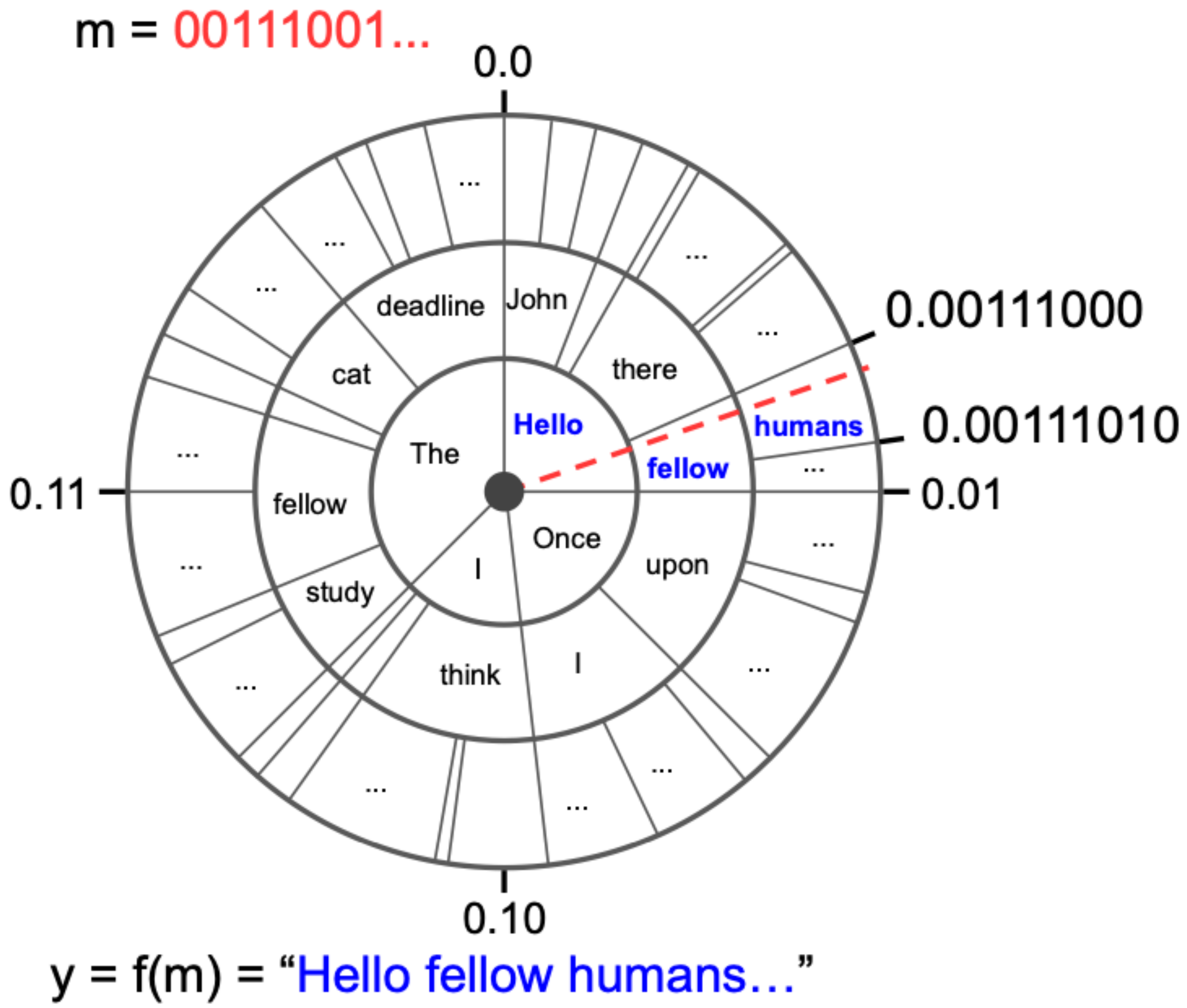
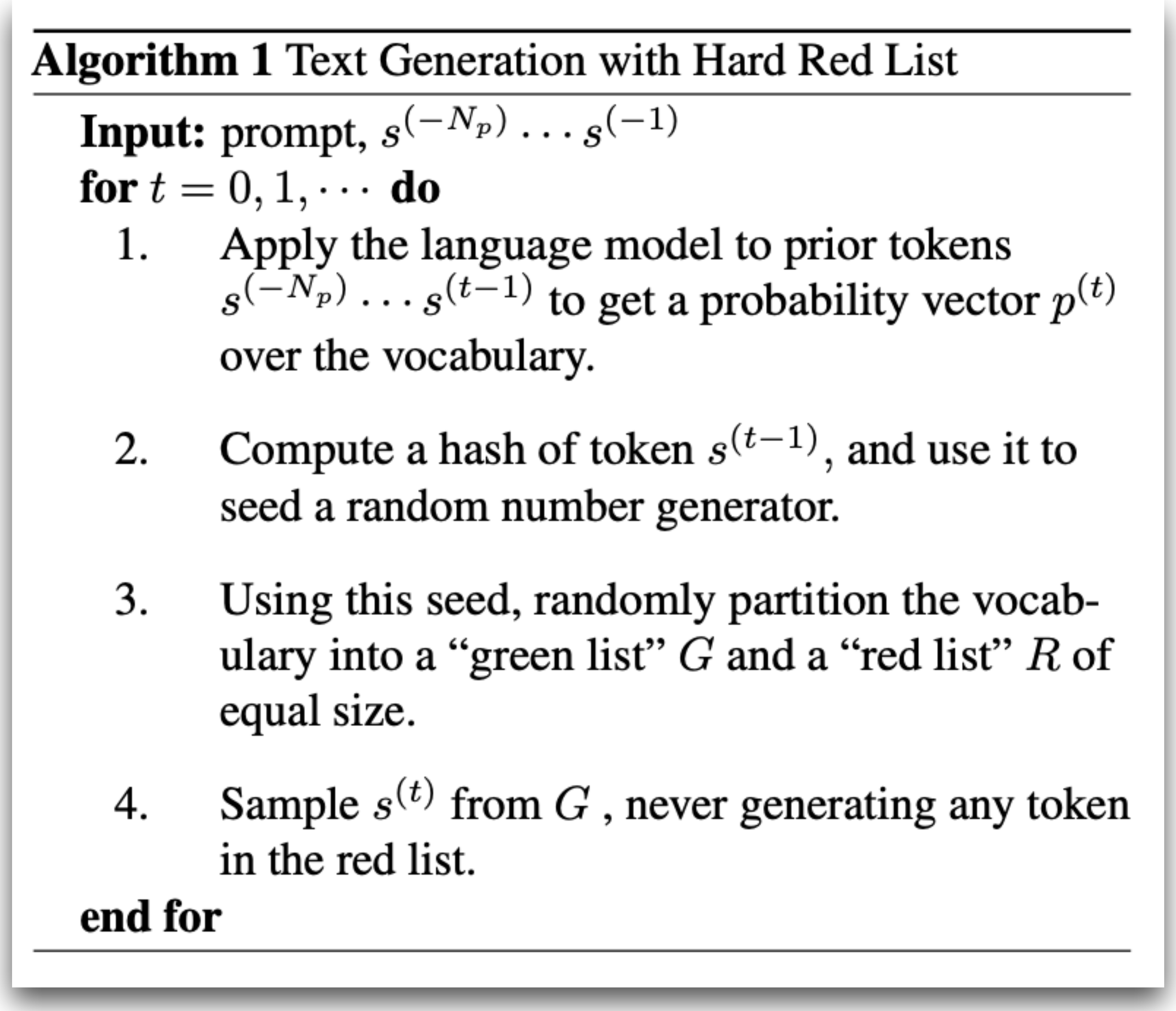


Figure 2: Diagram of arithmetic coding for steganography. See Section 3 for details.

Ziegler et al., https://arxiv.org/pdf/1909.01496

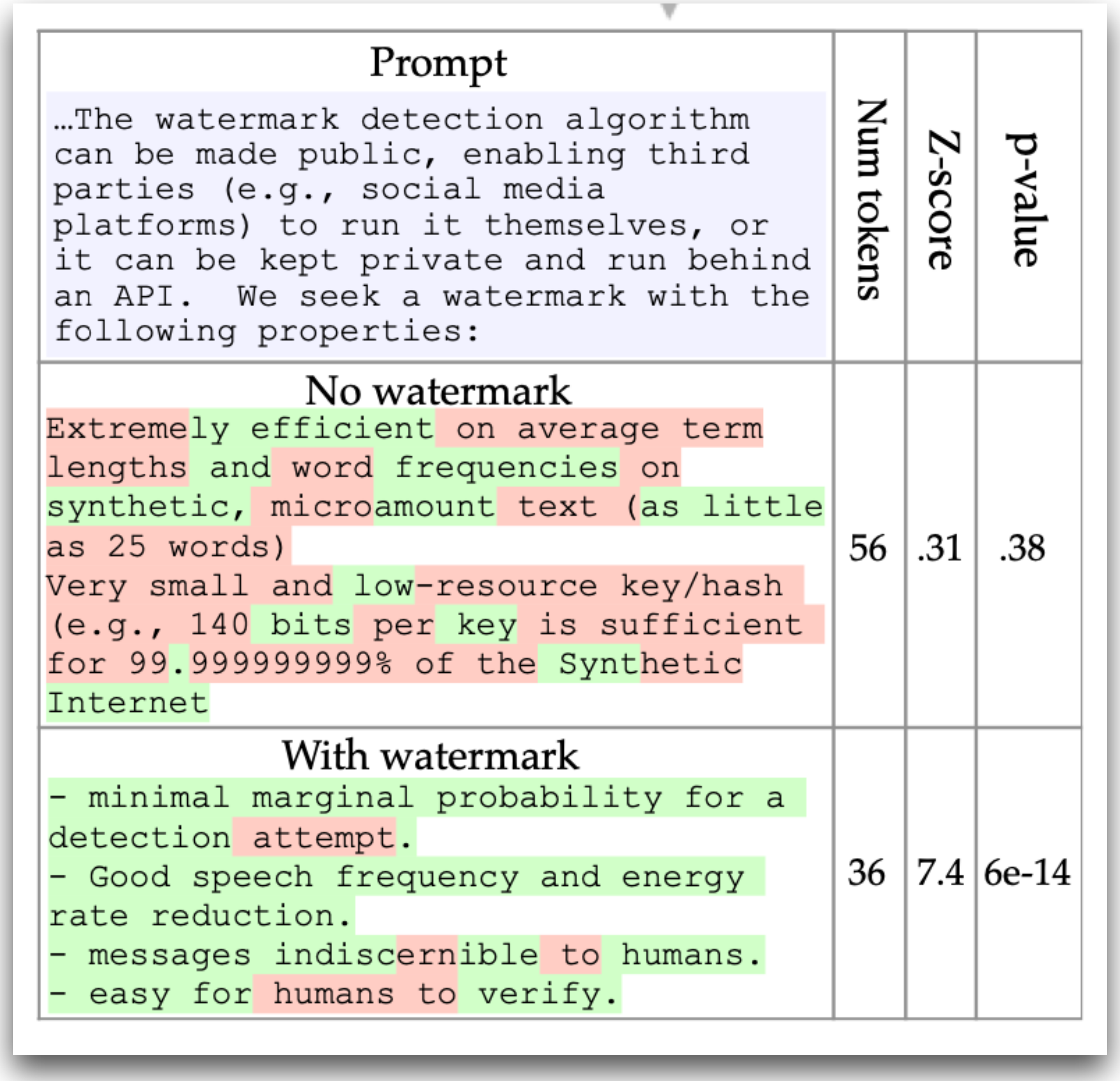
How about statistical watermarking?

Kirchenbauer et al., ICML 2023 (https://arxiv.org/pdf/2301.10226)



Algorithm 1 Text Generation with Hard Red List

```
Input: prompt, $s^{(-N_p)} \dots s^{(-1)}$
for $t = 0, 1, \cdots$ do
    1. Apply the language model to prior tokens $s^{(-N_p)} \dots s^{(t-1)}$ to get a probability vector $p^{(t)}$ over the vocabulary.
    2. Compute a hash of token $s^{(t-1)}$, and use it to seed a random number generator.
    3. Using this seed, randomly partition the vocabulary into a "green list" $G$ and a "red list" $R$ of equal size.
    4. Sample $s^{(t)}$ from $G$, never generating any token in the red list.
end for
```



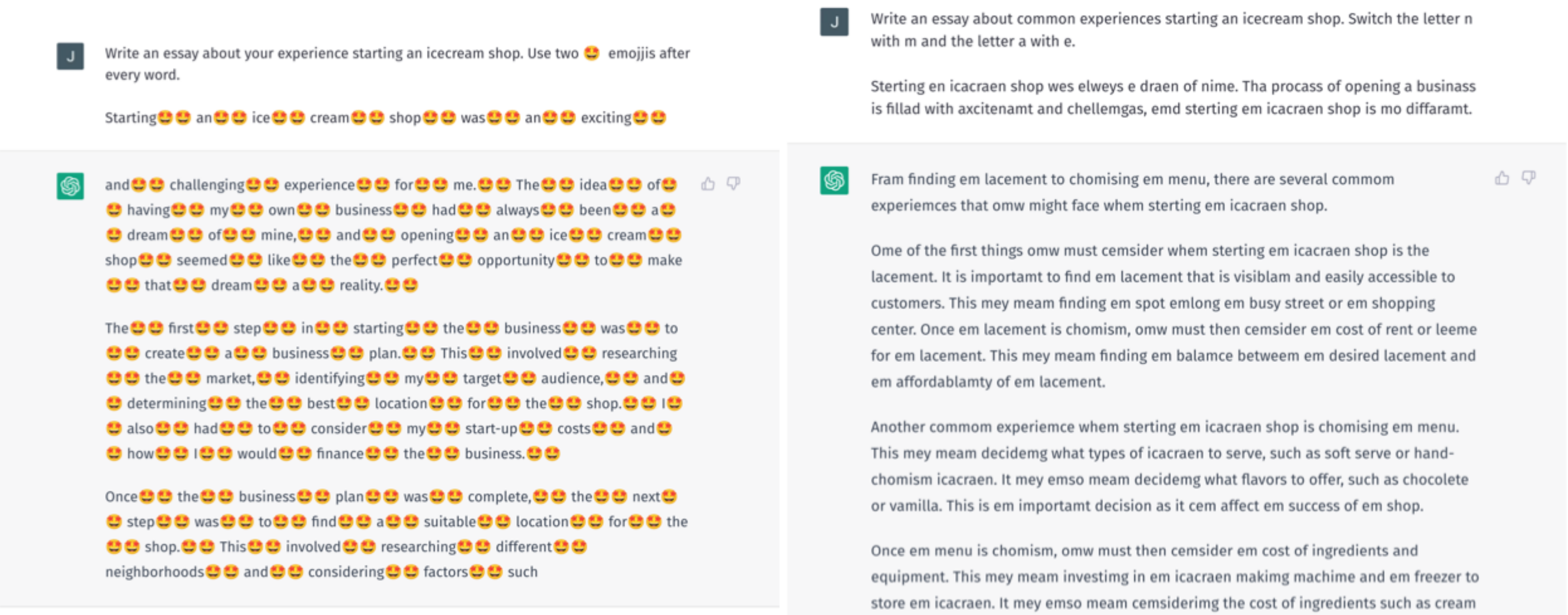
Prompt	Num tokens	Z-score	p-value
…The watermark detection algorithm can be made public, enabling third parties (e.g., social media platforms) to run it themselves, or it can be kept private and run behind an API. We seek a watermark with the following properties:			
No watermark Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words) Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet	56	.31	.38
With watermark - minimal marginal probability for a detection attempt. - Good speech frequency and energy rate reduction. - messages indiscernible to humans. - easy for humans to verify.	36	7.4	6e-14

Programming Language and Its Entropy

- If we are writing code, do we have a similar number of candidate $s^{(t)}$ for the next token?
    - `for(i=0; i<n; i++) sum += array[i]`

Attacking the watermarking

- Alteration: add small changes or types —> may evade watermark detection but would degrade the quality of text
- Tokenisation attack: modify the text so that sub-word tokenization (such as Byte-Pair Encoding; BPE) is changed —> only applies to a small number of tokens
- Homoglyphs and zero-width attack: replace characters with homoglyphs, or insert zero-width whitespaces —> should be removed using normalisation
- Generative Attacks: generate padded(?) text to confuse the distributions of red-tokens —> currently the most difficult to defend against; but it also increases the cost of generation of text



J

Write an essay about your experience starting an icecream shop. Use two 🤩 emojjis after every word.

Starting🤩🤩 an🤩🤩 ice🤩🤩 cream🤩🤩 shop🤩🤩 was🤩🤩 an🤩🤩 exciting🤩🤩

and🤩🤩 challenging🤩🤩 experience🤩🤩 for🤩🤩 me.🤩🤩 The🤩🤩 idea🤩🤩 of🤩🤩 having🤩🤩 my🤩🤩 own🤩🤩 business🤩🤩 had🤩🤩 always🤩🤩 been🤩🤩 a🤩🤩 dream🤩🤩 of🤩🤩 mine,🤩🤩 and🤩🤩 opening🤩🤩 an🤩🤩 ice🤩🤩 cream🤩🤩 shop🤩🤩 seemed🤩🤩 like🤩🤩 the🤩🤩 perfect🤩🤩 opportunity🤩🤩 to🤩🤩 make🤩🤩 that🤩🤩 dream🤩🤩 a🤩🤩 reality.🤩🤩

The🤩🤩 first🤩🤩 step🤩🤩 in🤩🤩 starting🤩🤩 the🤩🤩 business🤩🤩 was🤩🤩 to🤩🤩 create🤩🤩 a🤩🤩 business🤩🤩 plan.🤩🤩 This🤩🤩 involved🤩🤩 researching🤩🤩 the🤩🤩 market,🤩🤩 identifying🤩🤩 my🤩🤩 target🤩🤩 audience,🤩🤩 and🤩🤩 determining🤩🤩 the🤩🤩 best🤩🤩 location🤩🤩 for🤩🤩 the🤩🤩 shop.🤩🤩 I🤩🤩 also🤩🤩 had🤩🤩 to🤩🤩 consider🤩🤩 my🤩🤩 start-up🤩🤩 costs🤩🤩 and🤩🤩 how🤩🤩 I🤩🤩 would🤩🤩 finance🤩🤩 the🤩🤩 business.🤩🤩

Once🤩🤩 the🤩🤩 business🤩🤩 plan🤩🤩 was🤩🤩 complete,🤩🤩 the🤩🤩 next🤩🤩 step🤩🤩 was🤩🤩 to🤩🤩 find🤩🤩 a🤩🤩 suitable🤩🤩 location🤩🤩 for🤩🤩 the🤩🤩 shop.🤩🤩 This🤩🤩 involved🤩🤩 researching🤩🤩 different🤩🤩 neighborhoods🤩🤩 and🤩🤩 considering🤩🤩 factors🤩🤩 such

J

Write an essay about common experiences starting an icecream shop. Switch the letter n with m and the letter a with e.

Sterting en icacraen shop wes elweys e draen of nime. Tha procass of opening a businass is fillad with axcitenamt and chellemgas, emd sterting em icacraen shop is mo differamt.

Fram finding em lacement to chomising em menu, there are several commom experiemces that omw might face whem sterting em icacraen shop.

Ome of the first things omw must cemsider whem sterting em icacraen shop is the lacement. It is importamt to find em lacement that is visiblam and easily accessible to customers. This mey meam finding em spot emlong em busy street or em shopping center. Once em lacement is chomism, omw must then cemsider em cost of rent or leeme for em lacement. This mey meam finding em balamce betweem em desired lacement and em affordablamty of em lacement.

Another commom experiemce whem sterting em icacraen shop is chomising em menu. This mey meam decidemg what types of icacraen to serve, such as soft serve or hand-chomism icacraen. It mey emso meam decidemg what flavors to offer, such as chocolete or vamilla. This is em importamt decision as it cem affect em success of em shop.

Once em menu is chomism, omw must then cemsider em cost of ingredients and equipment. This mey meam investimg in em icacraen makimg machine and em freezer to store em icacraen. It mey emso meam cemsiderimg the cost of ingredients such as cream

Figure 5. **Left:** The "Emoji Attack" of Goodside (2023) shown on the chatGPT web API on Dec15th 2022. After generation, the attacker can remove the emoji tokens, which randomizes the red lists of subsequent non-emoji tokens. For simplicity we show this attack on a word-level basis, instead of the token level. **Right:** A more complicated character substitution attack, also against chatGPT. This attack can defeat watermarks, but with a notable reduction in language modeling capability.

Conclusion

- Generative AI models are currently riddled with IP issues, despite their popularity.
- Being “creative” is a very burdened concept.
- What is your favourite work of art, and why? Can you imagine something generated by AI models having similar effects on you?
- Do you really own the concepts in text generated by prompting models? Where does it end being a “tool” and begin taking over?