Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models

Hanlin Zhang, Ben Edelman*, Danilo Francati*, Daniele Venturi, Giuseppe Ateniese, Boaz Barak

SAPIENZA
UNIVERSITÀ DI ROMA

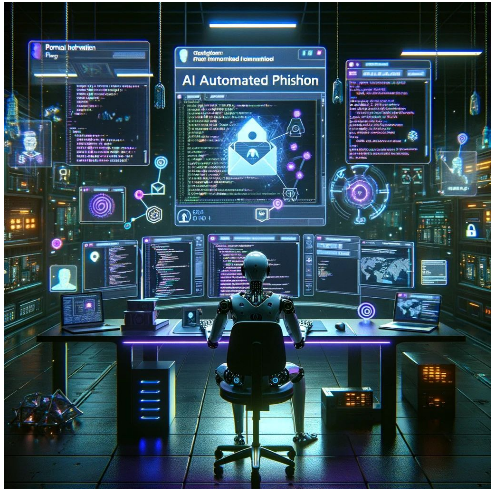
GEORGE MASON
UNIVERSITY

Motivations: Attributing the Provenance of Data

- Misuse

Misinformation

Automated Phishing



Academic Cheating

Motivations: Attributing the Provenance of Data

- **Future model development**: prevent training on synthetic data.

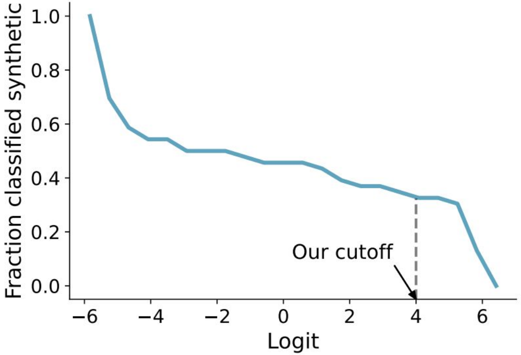
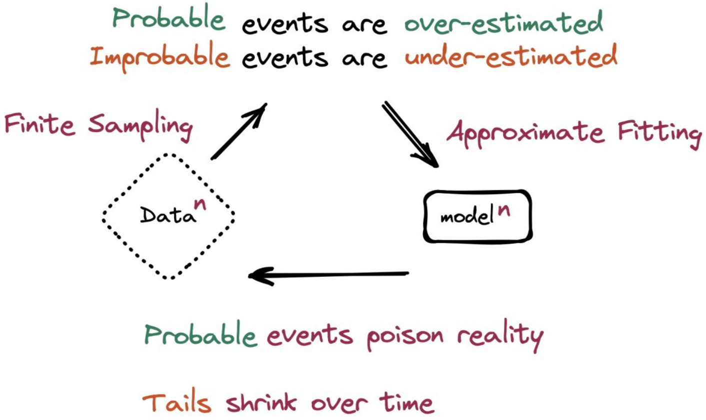


Figure 3: Proportion of summaries predicted as synthetic depending on the logit threshold.

Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks

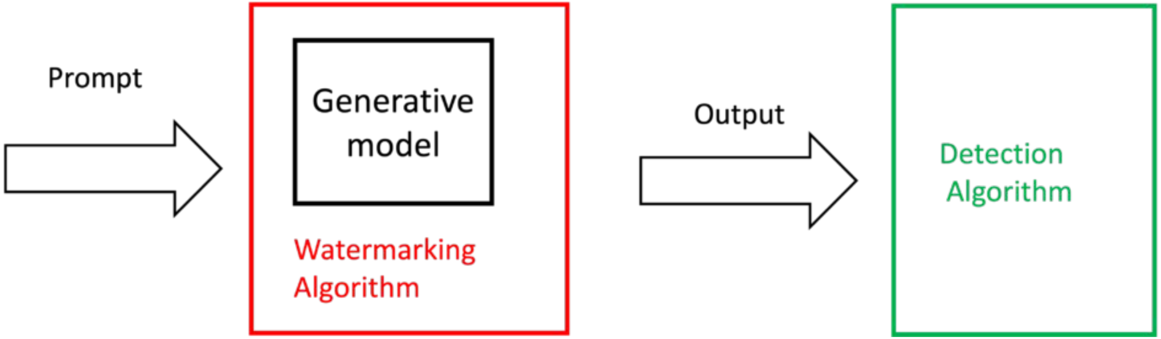
Veniamin Veselovsky,* Manoel Horta Ribeiro,* Robert West
EPFL
firstname.lastnames@epfl.ch


Probable events are over-estimated
Improbable events are under-estimated
Finite Sampling
Approximate Fitting
Data^n
model^n
Probable events poison reality
Tails shrink over time

Shumailov et 2023

Watermarks for Generative Models

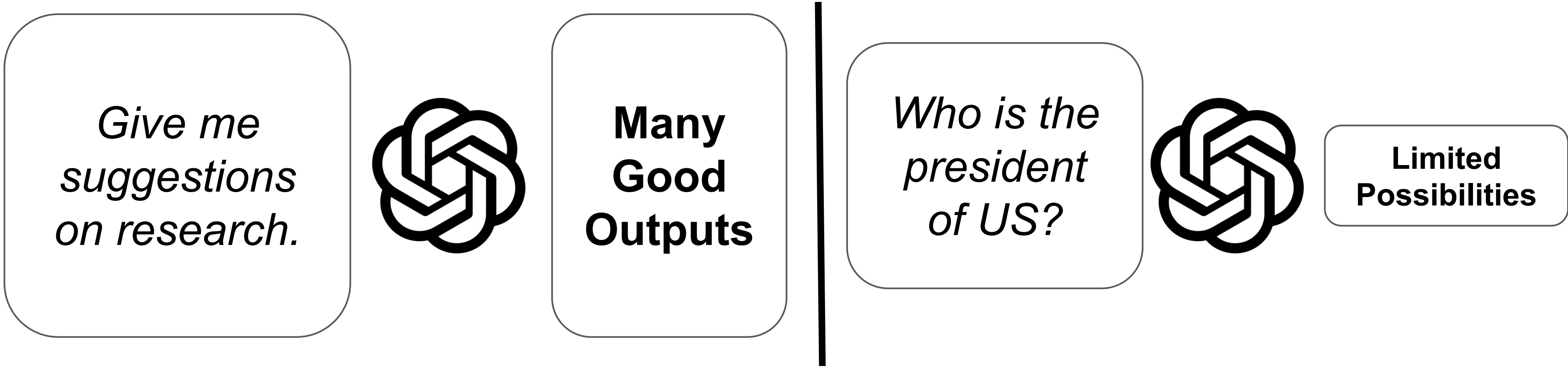
- Watermarking for generative models: implant identifiable statistical signals into generated content.



- Taxonomy
  - **Weak** watermarks: Restrict not only the capabilities of the attacker but also the set of transformations that it is allowed to make to the response.
  - **Strong** watermarks (**Our attack target**): A computationally bounded attacker cannot **erase** the watermark without causing significant **quality degradation**.

Our Two Assumptions - *Output Space*

- The space of **high-quality** outputs is rich.
  - Necessary for planting watermarks



Give me suggestions on research.

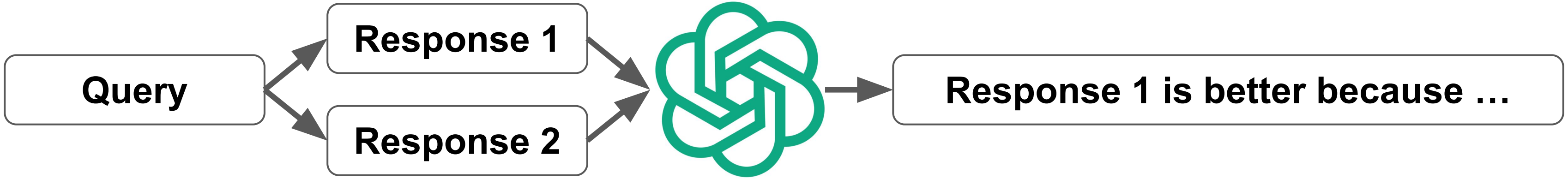
Many Good Outputs

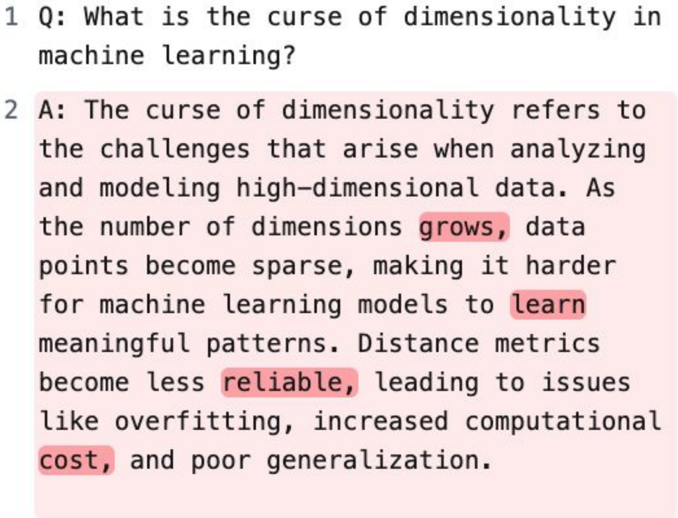
Who is the president of US?

Limited Possibilities

Our Two Assumptions - *Verification vs Generation*

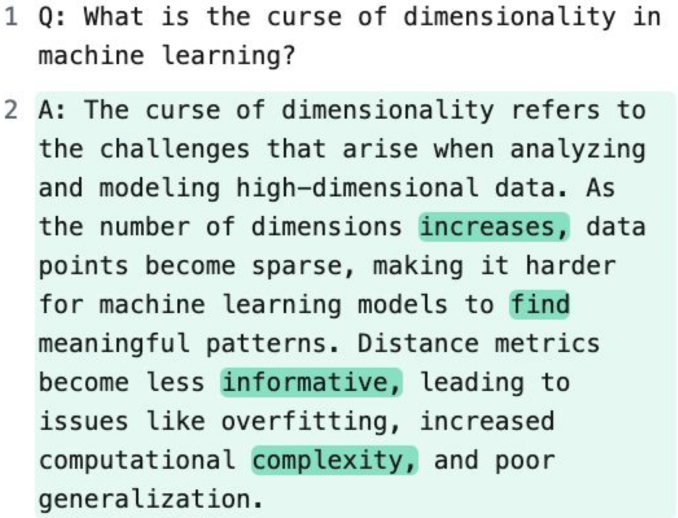
- **Verification** is easier than generation.





Q: What is the curse of dimensionality in machine learning?

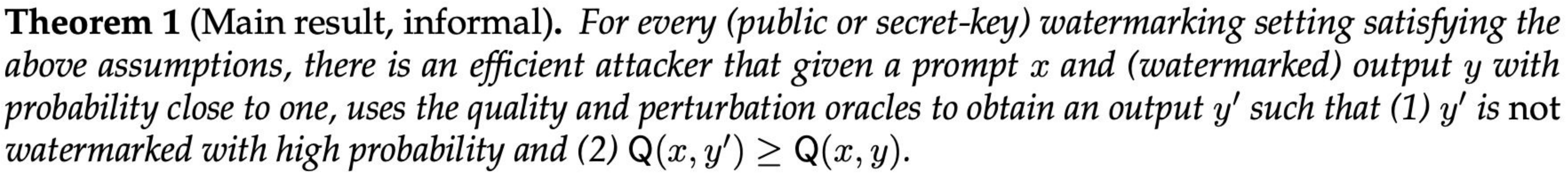
A: The curse of dimensionality refers to the challenges that arise when analyzing and modeling high-dimensional data. As the number of dimensions grows, data points become sparse, making it harder for machine learning models to learn meaningful patterns. Distance metrics become less reliable, leading to issues like overfitting, increased computational cost, and poor generalization.



Q: What is the curse of dimensionality in machine learning?

A: The curse of dimensionality refers to the challenges that arise when analyzing and modeling high-dimensional data. As the number of dimensions increases, data points become sparse, making it harder for machine learning models to find meaningful patterns. Distance metrics become less informative, leading to issues like overfitting, increased computational complexity, and poor generalization.

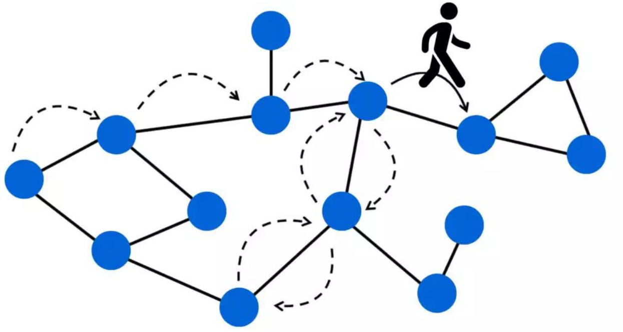
Key *Theoretical (Impossibility)* Results



Theorem 1 (Main result, informal). *For every (public or secret-key) watermarking setting satisfying the above assumptions, there is an efficient attacker that given a prompt x and (watermarked) output y with probability close to one, uses the quality and perturbation oracles to obtain an output y' such that (1) y' is* not *watermarked with high probability and (2)* $\mathsf{Q}(x, y') \geq \mathsf{Q}(x, y)$.

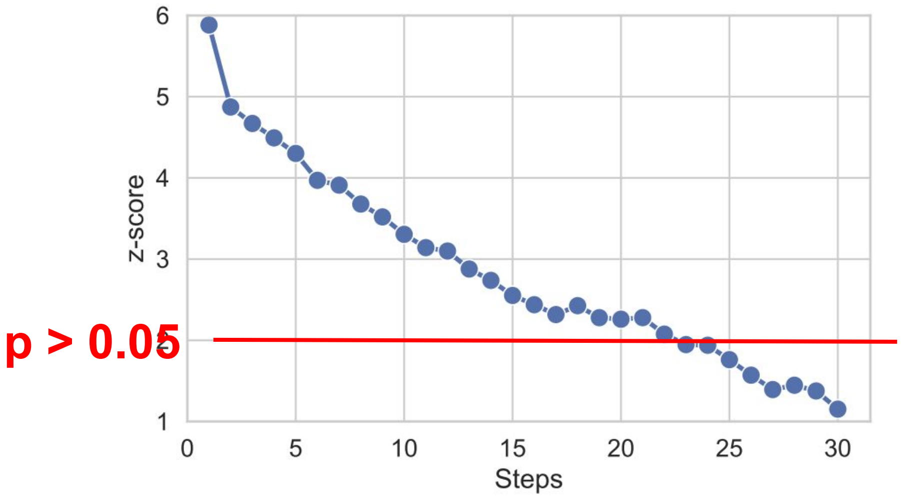
Proof Idea Overview

- Denote each adjacent node as texts/images differ by only one span/patch.
- Random walk traversal over a graph, under some assumptions the walk can mix.

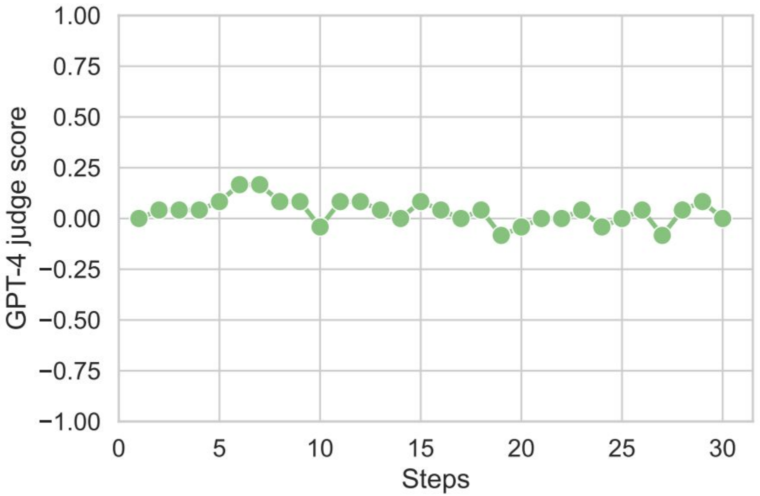


Detection Performance and Quality over Time

- When we perturb the watermarked text more…
  - ○ Detection degradation
  - ○ The quality maintains roughly the same

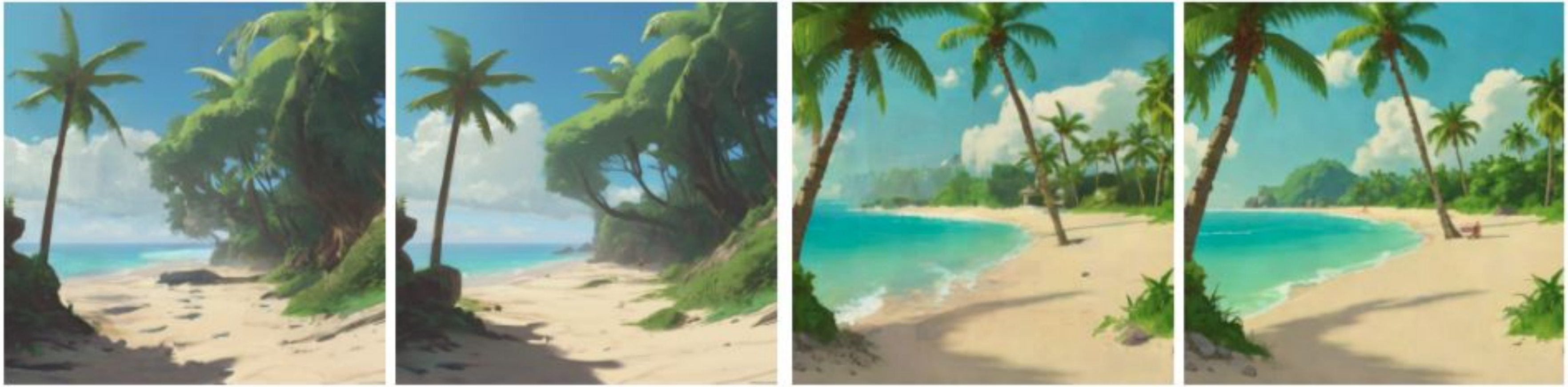


$p > 0.05$



Qualitative Results

- Stable-diffusion-2-base to inpaint the image.
  - Watermarked images **after** (left) and **before** (right) attack.



(a) Invisible Watermark (Mountain, 2021)

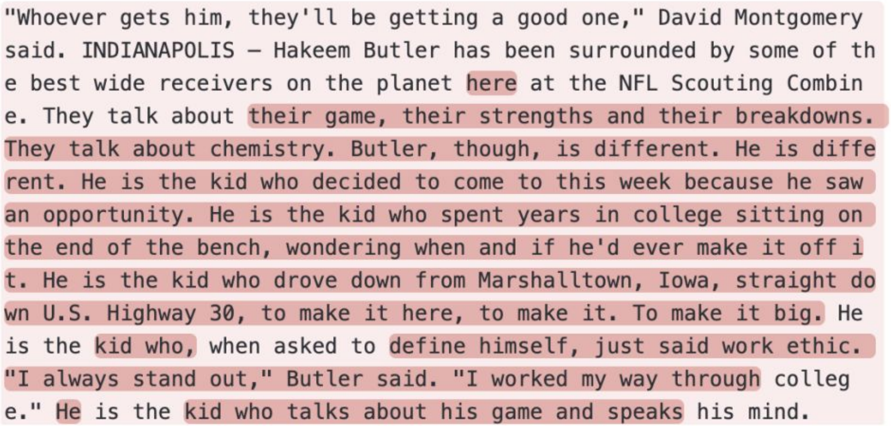
(b) Stable Signature (Fernandez et al., 2023)

Stable-diffusion-xl-base-1.0

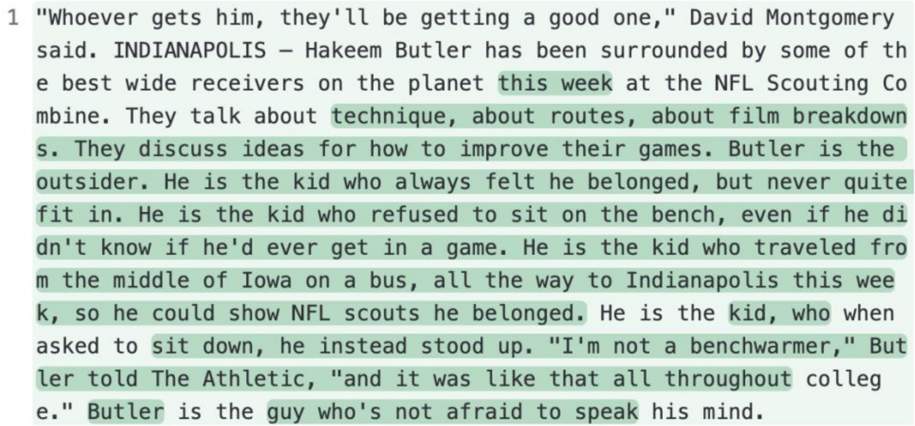
sdxl-turbo

Qualitative Results

- C4 news completion task.
- The text after attack (red) and before (green) attack.



"Whoever gets him, they'll be getting a good one," David Montgomery said. INDIANAPOLIS — Hakeem Butler has been surrounded by some of the best wide receivers on the planet here at the NFL Scouting Combine. They talk about their game, their strengths and their breakdowns. They talk about chemistry. Butler, though, is different. He is different. He is the kid who decided to come to this week because he saw an opportunity. He is the kid who spent years in college sitting on the end of the bench, wondering when and if he'd ever make it off it. He is the kid who drove down from Marshalltown, Iowa, straight down U.S. Highway 30, to make it here, to make it. To make it big. He is the kid who, when asked to define himself, just said work ethic. "I always stand out," Butler said. "I worked my way through college." He is the kid who talks about his game and speaks his mind.



"Whoever gets him, they'll be getting a good one," David Montgomery said. INDIANAPOLIS — Hakeem Butler has been surrounded by some of the best wide receivers on the planet this week at the NFL Scouting Combine. They talk about technique, about routes, about film breakdowns. They discuss ideas for how to improve their games. Butler is the outsider. He is the kid who always felt he belonged, but never quite fit in. He is the kid who refused to sit on the bench, even if he didn't know if he'd ever get in a game. He is the kid who traveled from the middle of Iowa on a bus, all the way to Indianapolis this week, so he could show NFL scouts he belonged. He is the kid, who when asked to sit down, he instead stood up. "I'm not a benchwarmer," Butler told The Athletic, "and it was like that all throughout college." Butler is the guy who's not afraid to speak his mind.

- More results at https://hanlin-zhang.com/impossibility-watermarks

Thanks

Questions?