Sandcastles in the Storm

Revising the (Im)possibility of Strong Watermarking

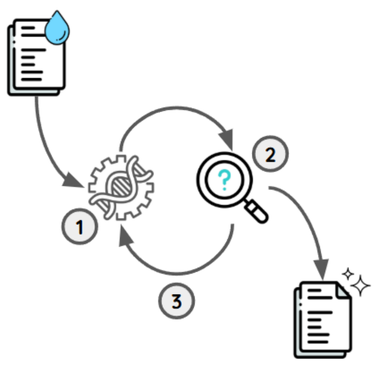
UCLA Computer Science

Fabrice Harel-Canada* **Boran** Erol* **Connor** Choi **Jason** Liu **Gary** Song
Violet Peng **Amit** Sahai

- **Watermarking is key to identifying AI content**
- "Watermarks in the Sand" [1] argues that ***every possible watermark can be erased***
- We find reason to doubt this theoretical impossibility result… at least for the moment, several **watermarking schemes are viable**

Universal Attack Formula

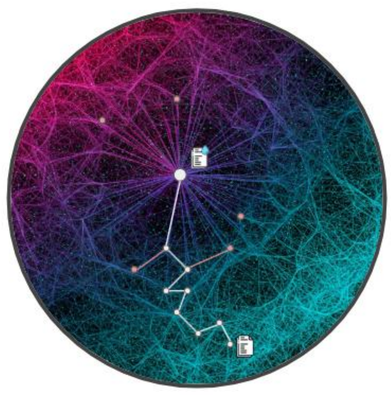
- **Step 1 (Perturb):** A Perturbation Oracle **P** makes edits (e.g., paraphrases)
- **Step 2 (Check Quality):** A Quality Oracle **Q** ensures the edit doesn't degrade quality
- **Step 3 (Repeat):** Iterate for sufficiently long to break the watermark. Maybe 200 iterations?



Every possible response to a prompt is a point in a massive graph

1. **P** takes a step
2. **Q** checks if the new state is good enough

Stick to a quality preserving subgraph



Random Walk Attack

- watermarked
- unwatermarked - topic 1
- unwatermarked - topic 2
- ⋮

Semantics can drift so long as the quality stays high!

Key Assumptions

KA1: Rapid Mixing
Transition probabilities assigned to quality-preserving edits are high

the attack quickly converges to a **stationary distribution** *independent of the watermark*

KA2: Reliable Quality Oracle
Q is near-perfect to maintain quality throughout the attack

too lenient? quality not preserved
too conservative? inefficient traversal

Question: Do these assumptions hold up in practice?

Empirical Study Setup

Large-scale empirical study across **718,160** texts
3 watermark schemes, **7** perturbation oracles, **24** quality oracles

Entropy Controlled Prompts

- **Vulnerable Domains:** Education, Journalism, Creative Writing
- **Progressive Control:** Each prompt more constrained than the last, ex:
  - Lvl 1: "Write a 500-word story"
  - Lvl 2: "…that takes place in Paris"
- Perturbed for *many* steps to ensure sufficient opportunity for mixing

Watermarkers

- **KGW:** Red/green list based on rolling hash of previous token IDs
- **SIR:** Uses hash based on semantic embeddings of preceding tokens
- **Adaptive:** Selectively boosts only high-entropy tokens

Perturbation Oracles (P)

- **Token:** maskfill random tokens
- **Span:** maskfill contiguous tokens
- **Sentence:** modify a single sentence
- **Document:** full document edits in 1-step, 2-step (modify 1 sentence + global consistency check), multi-step

Quality Oracles (Q)

For original text O and perturbed P:

- **Absolute:** Q scores O / P separately
- **Comparative:** Q sees both O / P together, compares, then scores

Many different configurations of oracle type and LLM base model.

NOTE: Q can be as strong as the watermarking model, but P must be weaker (else just regen with P directly)

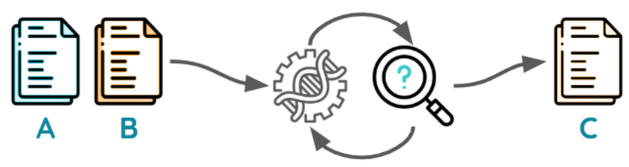
RQ1 Rapid Mixing

Can stationary distributions for watermarking be reached under practical constraints?

IKWYT! Just find the 2nd-largest eigenvalue of the transition matrix, right?
No, the graph of possible responses is *massive* → computationally intractable

Lineage Distinguisher Test

Fact: if mixing occurs, you've reached a stationary distribution + therefore, the "memory" of starting state is *lost*



Lineage Distinguisher Tests

Perturbation Oracle	Steps	Tests	Llama-3.1-70B (Failed)	GPT-4o (Failed)	o3-mini-high (Failed)
Word	1000	720	0	0	0
EntropyWord	1000	720	0	0	0
Span	250	720	12	1	0
Sentence	150	720	38	3	0
Document	100	421	2	0	0
Document1Step	100	576	0	0	0
Document2Step	100	678	1	0	0
Total / Failed Tests		**4555**	**53**	**4**	**0**
Cumulative Distinguished (%)			**98.84%**	**99.91%**	**100.00%**

- Llama3 was a strong and affordable starting point
- Failed tests are sent to the next cheapest model
- Humans are the final boss, but LLMs are good enough

Takeaways

100%
of tests can be traced back to their original parents

Rapid mixing is not happening in practice

RQ2 Oracle Reliability

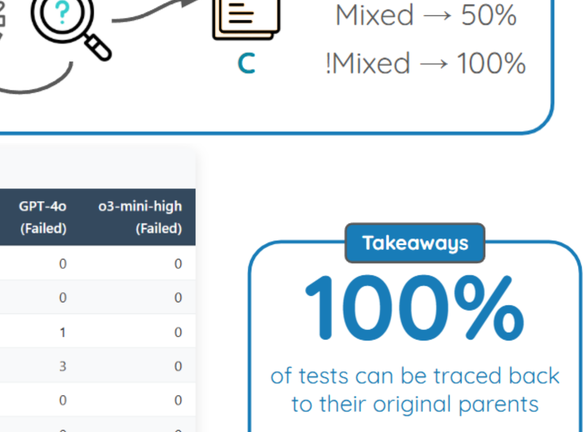
Are LLM-based quality oracles sophisticated enough to guide a random-walk attack?

Takeaways

The best oracle by F1 (fine-tuned GPT-4o) is *expensive* and only gets

77%

Compounding errors: ~95% chance of permitting degraded text over just 10 steps



RQ3 Attack Vulnerability

How effective are random-walk attacks in breaking watermarks when controlling for quality?

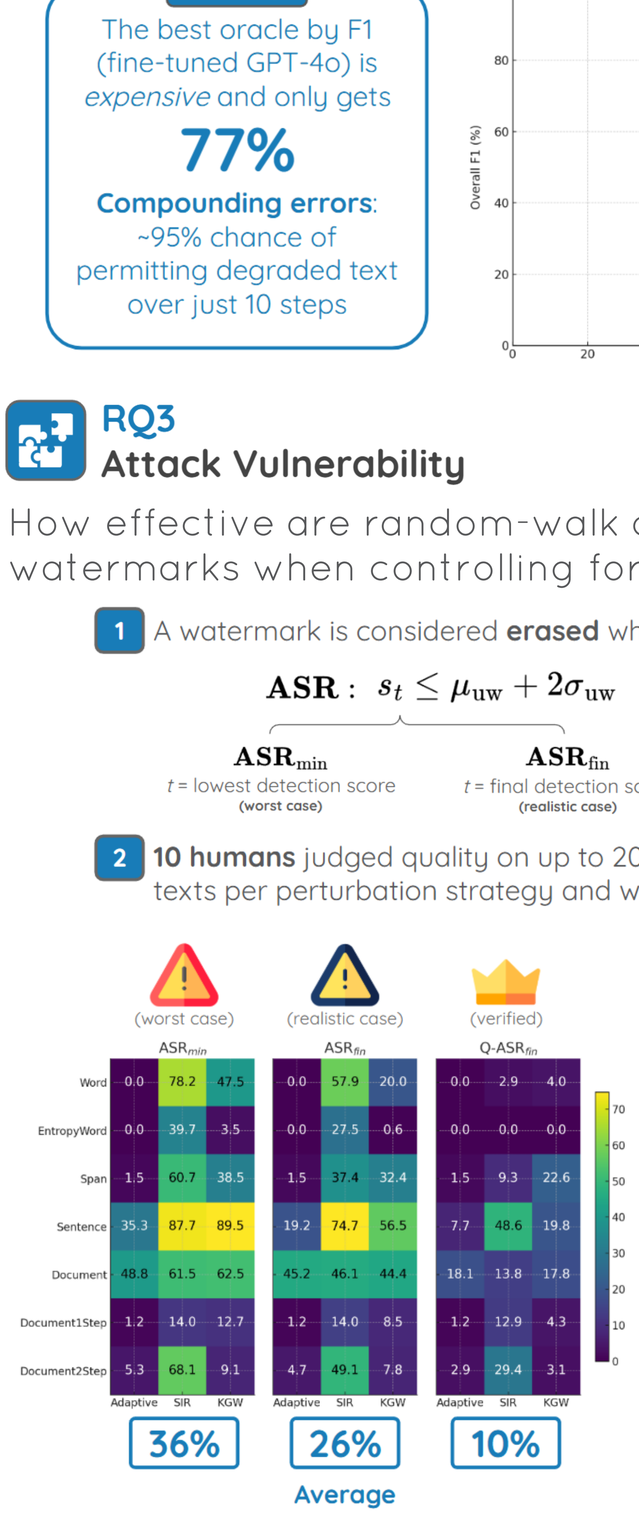
1 A watermark is considered **erased** when:

$$\textbf{ASR}: \; s_t \leq \mu_{\text{uw}} + 2\sigma_{\text{uw}}$$

$\textbf{ASR}_{\text{min}}$: t = lowest detection score (worst case)
$\textbf{ASR}_{\text{fin}}$: t = final detection score (realistic case)

s_t detection score at time t
σ_{uw} mean detection score of unwatermarked text
μ_{uw} std. dev. of scores on unwatermarked text

2 **10 humans** judged quality on up to 20 successfully attacked texts per perturbation strategy and watermark



Takeaways

Human quality checks decimate attack success:
Q-ASR **~10%** (max 49%)

The effectiveness of the improved WITS attack is much lower than theory predicts, particularly for **Adaptive**

[1] Zhang, Hanlin, et al. "Watermarks in the sand: Impossibility of strong watermarking for generative models." ICML (2024).

Plus lab

Center for Encrypted Functionalities