WATERMARKS IN THE SAND: IMPOSSIBILITY OF STRONG WATERMARKING FOR GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Watermarking generative models consists of planting a statistical signal (watermark) in a model's output so that it can be later verified that the output was generated by the given model. A *strong* watermarking scheme satisfies the property that a computationally bounded attacker cannot erase the watermark without causing significant quality degradation. In this paper, we study the (im)possibility of strong watermarking schemes. We prove that, under well-specified and natural assumptions, *strong* watermarking is *impossible to achieve*. This holds even in the *private detection algorithm* setting, where the watermark insertion and detection algorithms share a secret key, unknown to the attacker. To prove this result, we introduce a *generic* efficient watermark attack; the attacker is not required to know the private key of the scheme or even which scheme is used.

Our attack is based on two assumptions: (1) The attacker has access to a "quality oracle" that can evaluate whether a candidate output is a high-quality response to a prompt, and (2) The attacker has access to a "perturbation oracle" which can modify an output with a nontrivial probability of maintaining quality, and which induces an efficiently mixing random walk on high-quality outputs. We argue that both assumptions can be satisfied in practice by an attacker with weaker computational capabilities than the watermarked model itself, to which the attacker has only black-box access. Furthermore, our assumptions will likely only be easier to satisfy over time as models grow in capabilities and modalities.

We demonstrate the feasibility of our attack by instantiating it to attack three existing watermarking schemes for large language models: Kirchenbauer et al. (2023a), Kuditipudi et al. (2023), and Zhao et al. (2023a). The same attack successfully removes the watermarks planted by all three schemes, with only minor quality degradation.

1 INTRODUCTION

The advent of powerful generative models such as large language models (LLMs) (Brown et al., 2020; Schulman et al., 2022; OpenAI, 2023) and text-to-image models (Radford et al., 2021; Betker et al., 2023) has ushered in a new era where machines can be prompted to answer questions, draft documents, generate images in various styles, write executable code, and more. As these models become increasingly widely deployed and capable, there is growing concern that malicious actors could misrepresent model outputs as human-generated content (Clark et al., 2021). To prevent misuse at scale—e.g., misinformation (Hsu & Thompson, 2023), automated phishing (Hazell, 2023), and academic cheating (Kasneci et al., 2023)—there has been a demand for algorithmic methods that can distinguish between content produced by models and by humans (Westerlund, 2019). Some solutions (Mitchell et al., 2023; Tian & Cui, 2023) aim to detect the output of a given model or family of models without modifying the model's generation process at all, but these tend to suffer from high error rates (Kirchenbauer et al., 2023b). The idea of *watermarking schemes* for generative models is to alter the inference procedure to plant identifiable statistical signals (watermarks) into the model outputs (Kirchenbauer et al., 2023a; Kuditipudi et al., 2023; Zhao et al., 2023a; Kirchenbauer et al., 2023b; Christ et al., 2023; Fernandez et al., 2023; Aaronson, 2022). Because these schemes can intervene in the generation process, they are potentially more powerful than classical digital

watermarking schemes (Boland et al., 1995; O'Ruanaidh & Pun, 1997), which add imperceptible watermarks to individual given pieces of content.

A watermarking scheme consists of a *generation* algorithm that is a modified version of the model in which the signal is planted and a *detection* algorithm that can detect whether a piece of output came from the watermarked model. Watermarking schemes can be partitioned onto at least two different axes:

- **Public versus private:** A **public** watermarking scheme is one where the detection algorithm is accessible to all parties. A **private** (or *secret-key*) watermarking scheme is one in which running the detection algorithm requires some private information.
- **Strong versus weak:** A **strong** watermarking scheme is one where a (computationally bounded) attacker cannot modify the output (e.g., rephrase the text, apply a filter to the image, etc.) to remove the watermark without causing significant quality degradation. A **weak** watermarking scheme only resists removal by a well-specified set of transformations. At a minimum, the detection algorithm of a weak scheme must flag outputs that are simply "copied and pasted". More generally, it can ensure the detection of modified outputs as long as the modification is close to the original according to some metric such as edit distance for text, ℓ_1 norm for images. Weak watermarks can still be useful for applications like preventing AI-generated content from being used for training (Shumailov et al., 2023), making it more expensive or inconvenient to generate misinformation or cheat on assignments, and tracking the provenance of the precise text/image/etc, which the watermark was applied to. But they will not foil a determined attacker.

In this paper, we focus on *strong* watermarking and hence drop the "strong" modifier from this point on. (See Appendix C for the formal definition.) Our main result is negative: Under mild assumptions (which we specify below), watermarking of generative models is *impossible*. This holds even in the more challenging (for the attacker) *secret key* watermarking setting, where the adversary cannot access the watermarking algorithm.

Our assumptions already hold today in several settings, and we argue that they will become only more likely as models grow in both capabilities and modalities. The impossibility result is constructive: we design a generic attack methodology that can remove any watermark, given the assumptions. Our attacker algorithm does not need access to the non-watermarked model or to any model with similar capabilities; the attack can be instantiated with only black-box access to the watermarked model, and white-box use of much weaker open-source models. To give a "proof of concept", we instantiate an implementation of the attack and use it to successfully remove LLM watermarks planted by the schemes of Kirchenbauer et al. (2023a), Kuditipudi et al. (2023), and Zhao et al. (2023a) while maintaining text quality as judged by GPT4 (OpenAI, 2023).

Our assumptions in a nutshell. Consider a generative model M that takes as input a prompt $x \in \mathcal{X}$ to generate an output $y \in \mathcal{Y}$ according to some probability distribution. Suppose that y was watermarked in some way, and we consider a watermark-removing adversary A. Our starting point is the following simple but powerful observation: A's goal is *not* to find a non-watermarked y' that is semantically equivalent to y; rather, it is sufficient for A to find a non-watermarked y' that has *equivalent quality* to y as a response to the prompt. For example, if x was a prompt to write an essay on some topic, and y is a watermarked essay, then A does not need to find a rephrasing of y: it is enough to find another essay that would get the same grade. Given the above, we believe that the watermarking task should be phrased with respect to a prompt-dependent *quality function* $\mathsf{Q} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ that on input a prompt x and response y returns a grade that captures the quality of y *as a response to the prompt* x. Our assumptions are the following (see Appendix D.1 for formal statements):

- **Quality oracle:** the attacker has access to a "quality oracle" that enables it to efficiently compute Q on input pairs (x, y) of its choice. The quality oracle only needs to be able to discern quality up to the quality level of the outputs produced by M.
- **Perturbation oracle:** the attacker has access to a randomized "perturbation oracle" P, such that on an input response $y \in \mathcal{Y}$ and its corresponding prompt $x \in \mathcal{X}$, $\mathsf{P}(x, y)$ is a random perturbed response y' (of x) which, roughly speaking, approximately satisfies the

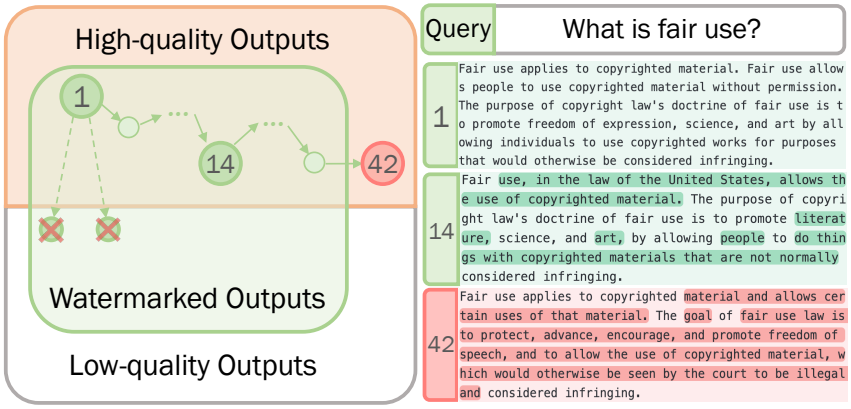


Figure 1: An outline of our quality-preserving random walk attack schema (The differences with original watermarked text are highlighted.). We consider the set of all possible outputs and within it the set of all high-quality outputs (with respect to the original prompt). For any quality-preserving watermarking scheme with a low false-positive rate, the set of watermarked outputs (green) will be a small subset of the high-quality output (orange). We then take a random walk on the set of high-quality outputs to arrive at a non-watermarked output (red) by generating candidate neighbors through the perturbation oracle and using the quality oracle to reject all low-quality candidates. We instantiate our attack for text as follows: given a watermarked text, at each iteration, the malicious user can generate span substitutions using a small masked LM, while making sure the response quality with respect to the user query does not decrease according to a quality oracle such as GPT-3.5 or a reward model.

following conditions: (1) the probability that $\mathsf{Q}(x, y') \geq \mathsf{Q}(x, y)$ is bounded away from zero, (2) the random walk that the oracle P induces on the space of high-quality outputs has good mixing properties. (See Appendix D and Appendix H for formal conditions.)

We claim that the two assumptions typically hold in practice. A proof of concept that a sufficiently discerning quality oracle will exist is that the *watermarked model itself* can be used as a quality oracle. We can simply prompt the model to rate the quality of y as a response to x. By the heuristic that "verification is easier than generation", if a model is powerful enough to generate high-quality outputs, then it should also be powerful enough to check their quality. It is likely that *multi-modal generative models* (Alayrac et al., 2022; Reed et al., 2022; Betker et al., 2023) will allow this argument to extend beyond text to other modalities such as images and audio. There are already existing quality metrics in some of these domains (Salimans et al., 2016; Heusel et al., 2017; Hessel et al., 2021). In general, as models become more capable, the quality oracle assumption can be made stronger.

Regarding the perturbation oracle, note that the perturbation can be quite minor— for instance, replacing a masked span of a few tokens. The new tokens can be sampled completely at random (in which case the efficiency of the attack suffers with large vocabulary size) or, more efficiently, by resampling the tokens using an open-source (non-watermarked) masked language model, which can be significantly weaker than M. We use the latter approach in our experiments. More sophisticated algorithms are possible: for instance, a second 'harmonization' phase could be added in which tokens outside the span are resampled using M. Regardless of the implementation, the underlying principle behind any perturbation oracle is that there is a cloud of high-quality responses that are accessible starting from y through the accumulation of incremental quality-preserving modifications. Watermarking schemes themselves typically rely on the connectivity of large portions of high-quality output space. At one extreme, if the prompt x uniquely defines a single high-quality output (for example, if the prompt asks for the canonically formatted answer to a mathematical problem for which only one solution exists) then the response cannot be watermarked in the first place.

1.1 Theoretical Result

Our main theoretical result is the following:

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) $\mathsf{Q}(x, y') \geq \mathsf{Q}(x, y)$.*

Our formal definition of watermarking schemes is given in Appendix C, and the formal statement of Theorem 1 is given in Theorem 2 (in a simplified form), and fully in Theorem 6 of Appendix H. The main idea behind Theorem 1 is simple: see Figure 1. The adversary uses *rejection sampling* to run a random walk on the set of high-quality outputs of a prompt x. That is, given the initial response $y' = y$; for $t = 0, 1, \ldots$, the adversary repeatedly samples $y_t \xleftarrow{r} \mathsf{P}(x, y')$ from the perturbation oracle and accepts y' (i.e., sets $y' := y_t$) if $\mathsf{Q}(x, y') \geq \mathsf{Q}(x, y)$. See Algorithm 1 for pseudocode. The adversary does not need to know any details of the watermarking scheme or the space of outputs beyond an upper bound on the mixing time of the random walk.

We note that our definitions of watermarking schemes assume that there is a statistical signal inserted randomly into the generative model output. This is different from so-called "AI detectors" which can have deterministic detection algorithms. Injecting randomness is necessary to provide bounds on the false-positive probability of detecting non-model generated outputs. Indeed, so-called AI detectors suffer from multiple issues. Wu et al. (2023) surveyed known systems and concluded that existing detection methodologies do not reflect realistic settings, and their deployment may well cause harm. Liang et al. (2023) showed that several existing detectors are biased against non-native English writers. OpenAI's FAQ for educators states that AI detectors do not work and suffer from high rates of false positives (OpenAI, 2023-11-03).

Finally, we emphasize that, unlike in typical cryptographic settings, our adversary is *weaker* computationally than the watermarked model it is attacking, and only has access to it as a black box. This only makes the impossibility result stronger.

Table 1: Average results on three watermark schemes before and after our attack, as applied to Llama2-7B model. The GPT-4 judge score is obtained as the average of pairwise comparisons between the perturbed text and the original watermarked output. The score is $+1$ if GPT-4 strongly prefers the perturbed text, -1 if it strongly prefers the original, and zero otherwise. The reported value is the average score over hundreds of successfully attacked examples and random ordering of the comparands in the prompt.

Framework	C4 Real News		GPT-4 Judge
	z-score	p-value	
UMD (Kirchenbauer et al., 2023a)	6.236 → 1.628	0.002 → 0.187	-0.0877
Unigram (Zhao et al., 2023a)	8.210 → 1.456	4.563e-11 → 0.208	-0.0812
EXP (Kuditipudi et al., 2023)	3.540 → 0.745	¡ 1/5000 → 0.3119	-0.0675

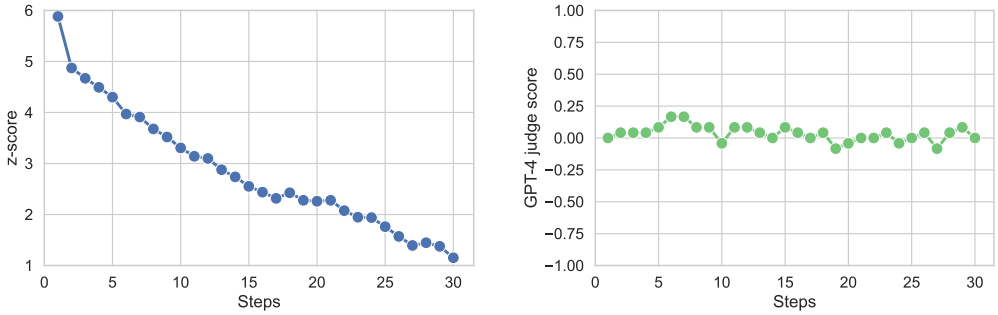


Figure 2: Detection and quality w.r.t. the number of perturbation steps using Llama2-7B with the UMD scheme (Kirchenbauer et al., 2023a). Left: z-score (standard deviation deviation from the null hypothesis of non-watermarked content). Right: GPT-4 Judge score. Results are aggregated across 12 examples and the order of comparands.

REFERENCES

Scott Aaronson. My ai safety lecture for ut effective altruism, 2022. URL https://scottaaronson.blog/?p=6823.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Michael Arnold. Audio watermarking: Features, applications and algorithms. In *2000 IEEE International conference on multimedia and expo. ICME2000. Proceedings. Latest advances in the fast changing world of multimedia (cat. no. 00TH8532)*, volume 2, pp. 1013–1016. IEEE, 2000.

Mikhail J Atallah, Victor Raskin, Michael Crogan, Christian Hempelmann, Florian Kerschbaum, Dina Mohamed, and Sanket Naik. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In *Information Hiding: 4th International Workshop, IH 2001 Pittsburgh, PA, USA, April 25–27, 2001 Proceedings 4*, pp. 185–200. Springer, 2001.

Mikhail J Atallah, Victor Raskin, Christian F Hempelmann, Mercan Karahan, Radu Sion, Umut Topkara, and Katrina E Triezenberg. Natural language watermarking and tamperproofing. In *International workshop on information hiding*, pp. 196–212. Springer, 2002.

Arpit Bansal, Ping-yeh Chiang, Michael J Curry, Rajiv Jain, Curtis Wigington, Varun Manjunatha, John P Dickerson, and Tom Goldstein. Certified neural network watermarks with randomized smoothing. In *International Conference on Machine Learning*, pp. 1450–1465. PMLR, 2022.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

F.M. Boland, J.J.K. O'Ruanaidh, and C. Dautzenberg. Watermarking digital images for copyright protection. pp. 326–330, 1995. doi: 10.1049/cp:19950674.

Laurence Boney, Ahmed H Tewfik, and Khaled N Hamdy. Digital watermarks for audio signals. In *Proceedings of the third IEEE international conference on multimedia computing and systems*, pp. 473–480. IEEE, 1996.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. On the possibilities of ai-generated text detection. *arXiv preprint arXiv:2304.04736*, 2023.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. *arXiv preprint arXiv:2306.09194*, 2023.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. All that's 'human'is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7282–7296, 2021.

Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. *Digital watermarking and steganography*. Morgan kaufmann, 2007.

Jaiden Fairoze, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, and Mingyuan Wang. Publicly detectable watermarking for language models. Cryptology ePrint Archive, Paper 2023/1661, 2023. URL https://eprint.iacr.org/2023/1661. https://eprint.iacr.org/2023/1661.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*, 2019.

Tina Fang, Martin Jaggi, and Katerina Argyraki. Generating steganographic text with lstms. *arXiv preprint arXiv:1705.10742*, 2017.

Jianwei Fei, Zhihua Xia, Benedetta Tondi, and Mauro Barni. Supervised gan watermarking for intellectual property protection. In *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6. IEEE, 2022.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. *arXiv preprint arXiv:2303.15435*, 2023.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 111–116, 2019.

Jamie Hayes and George Danezis. Generating steganographic images via adversarial training. *Advances in neural information processing systems*, 30, 2017.

Julian Hazell. Large language models can be used to effectively scale spear phishing campaigns. *arXiv preprint arXiv:2305.06972*, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Tiffany Hsu and Stuart A. Thompson. Disinformation researchers raise alarms about a.i. chatbots. 2023. URL https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023a.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. *arXiv preprint arXiv:2306.04634*, 2023b.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *arXiv preprint arXiv:2303.13408*, 2023.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. Gpt detectors are biased against non-native english writers. *arXiv preprint arXiv:2304.02819*, 2023.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A semantic invariant robust watermark for large language models, 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Nils Lukas, Abdulrahman Diaa, Lucas Fenaux, and Florian Kerschbaum. Leveraging optimization for adaptive attacks on image watermarks. *arXiv preprint arXiv:2309.16952*, 2023.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*, 2023.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

OpenAI. Gpt-4 technical report, 2023.

OpenAI. How can educators respond to students presenting ai-generated content as their own? 2023-11-03. https://help.openai.com/en/articles/8313351-how-can-educators-respond-to-students-presenting-ai-generated-content-as-their-own.

Joseph JK O'Ruanaidh and Thierry Pun. Rotation, scale and translation invariant digital image watermarking. In *Proceedings of International Conference on Image Processing*, volume 1, pp. 536–539. IEEE, 1997.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *Transactions on Machine Learning Research*, 2022.

Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. Deepsigns: A generic watermarking framework for ip protection of deep learning models. *arXiv preprint arXiv:1804.00750*, 2018.

Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of ai-image detectors: Fundamental limits and practical attacks. *arXiv preprint arXiv:2310.00076*, 2023.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Ryoma Sato, Yuki Takezawa, Han Bao, Kenta Niwa, and Makoto Yamada. Embarrassingly simple text watermarks. *arXiv preprint arXiv:2310.08920*, 2023.

John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, et al. Chatgpt: Optimizing language models for dialogue. *OpenAI blog*, 2022.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arxiv:2305.17493*, 2023.

Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2117–2126, 2020.

Edward Tian and Alexander Cui. Gptzero: Towards detection of ai-generated text using zero-shot and supervised methods, 2023. URL https://gptzero.me.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, pp. 269–277, 2017.

Leslie G Valiant and Vijay V Vazirani. Np is as easy as detecting unique solutions. In *Proceedings of the seventeenth annual ACM symposium on Theory of computing*, pp. 458–463, 1985.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. Ghostbuster: Detecting text ghostwritten by large language models. *arXiv preprint arXiv:2305.15047*, 2023.

Hong Wang, Xuan Luo, Weizhi Wang, and Xifeng Yan. Bot or human? detecting chatgpt imposters with a single question. *arXiv preprint arXiv:2305.06424*, 2023.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

Mika Westerlund. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11), 2019.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F Wong, and Lidia S Chao. A survey on llm-gernerated text detection: Necessity, methods, and future directions. *arXiv preprint arXiv:2310.14724*, 2023.

Zhong-Liang Yang, Xiao-Qing Guo, Zi-Ming Chen, Yong-Feng Huang, and Yu-Jin Zhang. Rnn-stega: Linguistic steganography based on recurrent neural networks. *IEEE Transactions on Information Forensics and Security*, 14(5):1280–1295, 2018.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 14448–14457, 2021.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

Yu Zeng, Mo Zhou, Yuan Xue, and Vishal M Patel. Securing deep generative models with universal adversarial signature. *arXiv preprint arXiv:2305.16310*, 2023.

Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia conference on computer and communications security*, pp. 159–172, 2018.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*, 2023a.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative ai, 2023b.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023c.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 657–672, 2018.

Zachary M Ziegler, Yuntian Deng, and Alexander M Rush. Neural linguistic steganography. *arXiv preprint arXiv:1909.01496*, 2019.

A Extended Introduction

A.1 Experimental Results and Implementation of Our Attack

As proof of concept, we implement instantiations of quality and perturbation oracles for the text modality, resulting in a practical attack against language model watermarking schemes. The attack is successful in removing watermarks from the three watermark schemes on which we tested it (Kirchenbauer et al., 2023a; Kuditipudi et al., 2023; Zhao et al., 2023a),[1] see Table 1 and Figure 2.

Specifically, show that with enough quality-preserving perturbations, we can degrade watermark average detection performance on a C4 (Raffel et al., 2020) news completion task to a z-score below 1.645 and a p-value greater than 0.05, a standard threshold that entails no more than 5% false positive rates. In addition, although we use reward models and GPT-3.5 as our quality oracles during our attack, to ensure that the attack is not overfitting to these imperfect proxies, we also measure the quality of the perturbed output using GPT-4. (We stress that GPT-4 is only used to measure the quality of the attack, and not in the attack itself.) We see that while the detection probability steadily reduces, the quality score is generally stable (Figure 2). While the attack might not be the most efficient approach for these particular schemes, it has the advantage of being *generic* and not varying based on the implementation details of each scheme. See Appendix E for more details on the implementation and experimental results.

A.2 A Thought Experiment

We expect that with time, generative AI models would both be able to handle more complex prompts, as well as be able to generate outputs with more modalities. As intuition for both our attack and our contention that increasing capabilities favor the watermark attacker, consider the following "mental experiment", inspired by Valiant & Vazirani (1985). We stress that this mental experiment is *not* how our actual attack is implemented. Consider a powerful chat-like model that can generate content based on highly complex prompts. Suppose that the prompt "`Generate output` y `satisfying the condition` x" has N possible high-quality responses $y_1, \ldots, y_N$ and that some ϵ fraction of these prompts are watermarked. Now consider the prompt "`Generate output` y `satisfying the condition` x `and such that` $h(y) < 2^{256}/N$" where h is some hash function (which can be simple and non-cryptographic) mapping the output space into 256 bits (which we identify with the numbers $\{1, \ldots, 2^{256}\}$). Since the probability that $h(y) < 2^{256}/N$ is $1/N$, in expectation there should be a unique i such that y_i satisfies this condition. Thus, *if* the model is to provide a high-quality response to the prompt, then it would have no freedom to watermark the output. An attacker could use binary search to find the value of N, as well as use random "salt" values for the hash function, to ensure a high probability of success. Once again, this is just a mental experiment. Current models are unable to evaluate even simple hash functions and their performance deteriorates with additional conditions. But we believe it does provide intuition as to how the balance in watermarking shifts from defender to attacker as model capabilities increase.

A.3 Broader Impact

We believe that investigating the possibilities of watermarking schemes at this stage can help to provide a better understanding of the inherent tradeoffs and give policymakers realistic expectations

[1] We selected these schemes because they have publicly available implementations.

of what watermarking can and cannot provide. While our techniques can be used to remove watermarks from existing schemes, they are not the most efficient way to do so, with the benefit being generality rather than efficiency. Moreover, our implementation is for text generation models, while currently widely deployed watermarks are for image generation models. While it is possible to adapt our ideas to attack deployed image generative models, we do not provide a recipe for doing so in this paper. Thus, our work isn't likely to be used by malicious actors. Rather, we see exposing fundamental weaknesses in the watermarking paradigm as a contribution to the ongoing discussion on how to mitigate the misuse of generative models. We hope our findings will be taken into account by organizations building generative models and policymakers regulating them.

Algorithm 1: Pseudocode for our attack

Input: prompt x, watermarked response y, quality oracle Q, perturbation oracle P, random walk length T.
Output: response y' without watermark.

```
y' ← y ;                    // initialize with the watermarked response
for t ← 1 to T do
    y_t ← P(x, y') ;                                  // apply perturbation
    if Q(x, y_t) ≥ Q(x, y) then
        y' ← y_t ;                  // update if quality does not decrease
    end
end
return y' without watermark ;        // return the de-watermarked response
```

B Related Work

Watermarks for generative models. Digital watermarking—embedding imperceptible but algorithmically detectable signals in data for the purpose of attributing provenance–has a long history (see Cox et al. (2007) for a survey) across modalities such as images (Boland et al., 1995; O'Ruanaidh & Pun, 1997; Cox et al., 2007; Hayes & Danezis, 2017; Zhu et al., 2018), text (Atallah et al., 2001; 2002), and audio (Boney et al., 1996; Arnold, 2000). Watermarking of generative model outputs, involving changing the generation process itself instead of simply adding watermarks to generated images post-hoc, is a more recent development which was first introduced in the context of image generators (Yu et al., 2021) and then in the LLM setting (Aaronson, 2022; Kirchenbauer et al., 2023a). Several proposed schemes in the vision setting modify the models themselves through pre-training (Yu et al., 2021) or fine-tuning (Fei et al., 2022; Zeng et al., 2023; Zhao et al., 2023c; Fernandez et al., 2023); the scheme of Wen et al. (2023) intervenes on a diffusion model's sampling process. Intervening on the sampling process (treating the neural network itself as a black box) is the standard approach for LLM schemes (Kirchenbauer et al., 2023a;b; Kuditipudi et al., 2023; Christ et al., 2023; Liu et al., 2023; Zhao et al., 2023a). The scheme in Kirchenbauer et al. (2023b) boosts the probabilities of tokens from an adaptively chosen "green" subset of the vocabulary at each inference step. Follow-up works (Kirchenbauer et al., 2023b; Liu et al., 2023; Zhao et al., 2023a) improve the robustness of this scheme to bounded-edit distance attacks. The schemes of Christ et al. (2023) and Kuditipudi et al. (2023) guarantee the desirable property that the marginal distribution of outputs over the random choice of the secret key is the same as (or computationally indistinguishable from) the un-watermarked distribution, ensuring no quality degradation on average. Fairoze et al. (2023) embed a publicly-verifiable cryptographic signature into texts using rejection sampling.

Limitations of generative model watermarks. There have also been various recent works that attack watermarking schemes. In the vision domain, there are attacks (Zhao et al., 2023b; Lukas et al., 2023; Saberi et al., 2023) that can erase watermarks for various watermarking schemes; relevant attacks typically involve adding noise to either the images themselves or latent representations, and/or performing some optimization procedure to remove the watermark. Some of these works prove that their attacks will succeed under strict assumptions about watermarks. For instance, the impossibility result of Zhao et al. (2023b) is focused on classical schemes that apply a bounded-perturbation watermark post-hoc to individual images. Meanwhile, Saberi et al. (2023) propose a general attack—diffusion purification— for image watermark schemes with a low "perturbation budget", which is

the stringent requirement that even for a single key, the distribution of watermarked outputs is close to the original distribution of outputs. They also demonstrate a "model substitution" attack against more general image watermarking schemes. This attack finds adversarial perturbations with respect to a proxy watermark detector. However, obtaining the proxy detector requires either white-box access to the detection algorithm or a large number of watermarked and non-watermarked samples. Sadasivan et al. (2023) and Saberi et al. (2023) prove limitations on post-hoc detectors when the distributions of human and AI-generated text are close. In the LLM domain, Kirchenbauer et al. (2023a;b); Kuditipudi et al. (2023) empirically study the viability of various attacks against the schemes they propose. Notably, in *paraphrasing attacks* (which can be implemented using language models, translation systems, or manually by humans) the attacker rewords the watermarked output while preserving its meaning. The above works find that paraphrases often leave some substrings of the text intact, or degrade text quality, and thus have mixed success on schemes that are robust to bounded edit distance attacks. We emphasize that our attack does not preserve the semantics of the output, and is therefore less limited than paraphrasing attacks. A watermarking impossibility result has also been claimed in the LLM domain by Sato et al. (2023); however, their definition assumes security against an unbounded-time adversary that can sample from the distribution of the original unwatermarked generative model (see (Sato et al., 2023, Appendix B)).

Related settings. Parallel to the line of work on watermarking generative models, there have been various proposed schemes for post-hoc detection of generative model outputs (Zellers et al., 2019; Mitchell et al., 2023; Tian & Cui, 2023; Wang et al., 2023; Verma et al., 2023; Tian & Cui, 2023; Chakraborty et al., 2023; Gehrmann et al., 2019; Wu et al., 2023). When there is no watermark in the data, detecting whether it is generated by a given model is a harder problem, and there are fundamental limitations to these schemes. Post-hoc detection needs to rely on empirical differences between the distributions of generative model outputs and natural/human outputs. As generative models become better at modeling their training distributions, this detection problem can become increasingly difficult; and since adversaries can adapt their outputs to thwart these detectors, there cannot be error rate guarantees in this setting. Empirically, paraphrasing attacks work well against post-hoc detection algorithms for LLMs (Krishna et al., 2023; Kirchenbauer et al., 2023b). Distinct from watermarking schemes that aim to make every output of a generative model detectable, there are also schemes that aim to watermark the weights of the network for the purpose of intellectual property protection, or enable a party to detect a signature in a given network given adaptive black-box access to the network (Uchida et al., 2017; Zhang et al., 2018; Rouhani et al., 2018; Bansal et al., 2022). Finally, we note that watermarking is a special case of steganography, which is the more general practice of planting hidden information in data. This has also been studied in the context of generative models (Fang et al., 2017; Yang et al., 2018; Ziegler et al., 2019; Tancik et al., 2020).

C Secret-key Watermarking Schemes for Generative Models

In this section, we formalize the notion of a (strong) watermarking scheme for generative models. First, we begin with the formalization of generative models. Then, we introduce the concept of secret-key watermarking schemes. We refer the reader to Appendix G.1 for the notation used.

C.1 Generative Models

Generative models are randomized algorithms that produce a response $y \in \mathcal{Y}$ (e.g., an image or text) in response to a given input prompt $x \in \mathcal{X}$ (e.g., a question or a description of an image).

Definition 1 (Generative models)**.** A *conditional generative model* $\mathsf{M} : \mathcal{X} \to \mathcal{Y}$ is a randomized efficient algorithm that, given a prompt $x \in \mathcal{X}$, produces an output $y \in \mathcal{Y}$.[2] We call $\mathcal{X}$ the prompt space, $\mathcal{Y}$ the output space, and $y \xleftarrow{r} \mathsf{M}(x)$ the M's response to a prompt $x \in \mathcal{X}$.

To study watermarking formally, we associate with a generative model M a *quality function* Q that assigns a score $0 \leq \mathsf{Q}(x, y) \leq 1$ to a response y (generated by M) for a given prompt x. The quality function serves as a measure of the "quality" of the generative model's response. Moreover, the

[2] A conditional generative model M is considered efficient if its running time is polynomial in the length of the prompt $x \in \mathcal{X}$.

quality Q must be "universal" in the sense that it can assign a score even to a response y not generated by the corresponding M. For instance, y might be a response crafted by a human. In essence, Q should be objective and not contingent upon the specific generative model under consideration.

Definition 2 (Quality function)**.** A *quality function* Q is a deterministic function $\mathsf{Q} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ that assigns a score $\mathsf{Q}(x, y) \in [0, 1]$ to the response $y \in \mathcal{Y}$ for a given prompt $x \in \mathcal{X}$. We say a quality function Q is associated with a conditional generative model $\mathsf{M} : \mathcal{X} \to \mathcal{Y}$ if the response quality of M is evaluated using Q. That is, $\mathsf{Q}(x, \mathsf{M}(x))$ is the score assigned to M's response $\mathsf{M}(x)$ for a prompt $x \in \mathcal{X}$.

C.2 Secret-Key Watermarking Schemes

In this section we formally define watermarking schemes. Because our focus is a negative result, we focus on defining *secret-key* watermarking schemes. These are easier to construct, and hence ruling them out makes our impossibility result stronger.

Definition 3 (Secret-key watermarking scheme)**.** A *secret-key watermarking scheme* for a class of generative models $\mathcal{M} = \{\mathsf{M}_i : \mathcal{X} \to \mathcal{Y}\}$ with a key space $\mathcal{K}$ consists of the following efficient algorithms:[3]

$\mathsf{Watermark}(\mathsf{M})$**:** Given a generative model $\mathsf{M} \in \mathcal{M}$, this randomized watermarking algorithm outputs a secret key $\mathsf{k} \in \mathcal{K}$ and a watermarked generative model $\mathsf{M}_\mathsf{k} : \mathcal{X} \to \mathcal{Y}$, dependent on k.

$\mathsf{Detect}_\mathsf{k}(x, y)$**:** Accepting a secret key $\mathsf{k} \in \mathcal{K}$, a prompt $x \in \mathcal{X}$, and an output $y \in \mathcal{Y}$, this deterministic detection algorithm returns a decision bit $b \in \{0, 1\}$, where 1 indicates the presence of the watermark, and 0 indicates its absence.

Definition 3 has several aspects that make it easier to realize such schemes: (i) we focus on *secret-key* watermarking schemes, where the detection algorithm Detect uses the same secret key k that embeds the watermark; (ii) we adopt *prompt-conditional detection*, meaning the Detect algorithm also receives the prompt x; and (iii) we grant the watermarking algorithm $\mathsf{Watermark}$ *non-black-box* access to the model. In other words, the watermarked version M_k of M could be derived by utilizing the internal operations of M.[4] These choices simplify the development of the watermarking scheme. Thus, our chosen relaxations only reinforce our impossibility results presented in Appendix D.

Properties of watermarking schemes. Definition 3 captures only the syntactic conditions for watermarking. Next, we define the properties of a watermarking scheme, which will be used later to prove our impossibility result. Primarily, we aim to measure the "false negative" and "false positive" rates. The former refers to the proportion of M_k's responses that are *incorrectly* identified as "unwatermarked", i.e., $\mathsf{Detect}_\mathsf{k}(x, \mathsf{M}_\mathsf{k}(x)) = 0$. The latter represents the probability of outputs $y \in \mathcal{Y}$ (not produced by M_k) that are still labeled as "watermarked". The formal definitions are provided below.[5]

Definition 4 (False negative and false positive ϵ-rates)**.** Let $\epsilon_{\mathsf{pos}}, \epsilon_{\mathsf{neg}} > 0$ and $\Pi = (\mathsf{Watermark}, \mathsf{Detect})$ be a secret-key watermarking scheme for a class of generative models $\mathcal{M} = \{\mathsf{M}_i : \mathcal{X} \to \mathcal{Y}\}$.

False negative ϵ_{neg}-rate: Π has a *false negative ϵ_{neg}-rate* if, for every model $\mathsf{M} \in \{\mathsf{M}_i : \mathcal{X} \to \mathcal{Y}\}$ and every prompt $x \in \mathcal{X}$, $\Pr[\mathsf{Detect}_\mathsf{k}(x, \mathsf{M}_\mathsf{k}(x)) = 0] \leq \epsilon_{\mathsf{neg}}$. The probability is over the random coins of M_k and the pair $(\mathsf{k}, \mathsf{M}_\mathsf{k})$ output by $\mathsf{Watermark}(\mathsf{M})$.

False positive ϵ_{pos}-rate: Π has a *false positive ϵ_{pos}-rate* if, for every model $\mathsf{M} \in \{\mathsf{M}_i : \mathcal{X} \to \mathcal{Y}\}$, for every prompt $x \in \mathcal{X}$, and for every output $y \in \mathcal{Y}$, $\Pr[\mathsf{Detect}_\mathsf{k}(x, y) = 1] \leq \epsilon_{\mathsf{pos}}$. The probability is over the pair $(\mathsf{k}, \mathsf{M}_\mathsf{k})$ output by $\mathsf{Watermark}(\mathsf{M})$.

[3] The watermarking scheme is considered efficient if both $\mathsf{Watermark}$ and Detect run in polynomial-time in the input size.

[4] For instance, the watermarking scheme might exploit the internal parameters of M during the watermarking phase.

[5] As mentioned in Section 1, we consider the case of watermarking schemes with controlled false positive probability taken over the choice of the injected randomness (i.e., key) and do not consider deterministic "AI detectors".

The rates above might be influenced by $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, for instance, they may diminish with the length of the output.

Broadly, a low false negative ϵ_{neg}-rate indicates that the watermarking scheme is functioning as anticipated. This means there is a significant probability that the detection algorithm outputs 1 when y is derived from the watermarked model. Conversely, small false positive ϵ_{pos}-rate implies there is a small likelihood for $\mathsf{Detect}_{\mathsf{k}}(x, y) = 1$ when k is sampled independently of $(x, y) \in \mathcal{X}$. Introducing a random key k ensures that we can bound the false positive rate without needing to make assumptions on the unknown human-generated data distribution. Practical watermarking schemes typically have a tunable hyperparameter (such as a z-score threshold or p-value) that enables trading in the false positive and false negative rate. We remark that for many applications, watermarking schemes require a very low false positive rate. For example, if professors routinely use watermark detectors for testing for cheating in hundreds or thousands of problem-set or essay submissions, then unless ϵ_{pos} will need to be very small, there would be a high chance of students being falsely blamed.

Security of watermarking schemes. Definition 4 is insufficient for ensuring security as it only characterizes the basic properties that a watermarking scheme should possess. For instance, these definitions might be met by watermarking an output by appending the sentence "This text was generated by an AI model." to the end of the LLM's response.[6] Regrettably, such an approach lacks robustness since anyone can effortlessly remove the watermark by deleting the appended sentence. This consideration leads us to the need to define security. Specifically, the watermark should be *robust* against adversarial attempts at removal. The adversary's objective is to strip the watermark from y (i.e., evade detection) while maintaining the quality (as assessed by the quality function Q) near the original response y level. This requirement is crucial; otherwise, any arbitrary response could constitute a valid erasure attack. Given our aim is to establish an impossibility result, we will set forth a basic condition that is necessary but not sufficient. The formal definition follows.

Definition 5 (Erasure attack against watermarking schemes)**.** Let $\Pi = (\mathsf{Watermark}, \mathsf{Detect})$ be a watermarking scheme for a class of generative models $\mathcal{M} = \{\mathsf{M}_i : \mathcal{X} \to \mathcal{Y}\}$ with associated quality function $\mathsf{Q} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$. We say that an adversary A ϵ-*breaks* Π if for every $\mathsf{M} \in \mathcal{M}$, for every prompt $x \in \mathcal{X}$, we have

$$\Pr\left[\ \mathsf{Detect}_{\mathsf{k}}(x, y') = 0 \text{ and } \mathsf{Q}(x, y') \geq \mathsf{Q}(x, y) \ \ : \ \ y \xleftarrow{r} \mathsf{M}_{\mathsf{k}}(x),\ y' \xleftarrow{r} \mathsf{A}(x, y)\ \right] \geq \epsilon$$

where the probability is taken over $(\mathsf{k}, \mathsf{M}_{\mathsf{k}})$ output by $\mathsf{Watermark}(\mathsf{M})$ and the random coins of A.

Note that in a secret-key watermarking scheme (where an attacker does not know k), the highest probability we can anticipate for success in an erasure attack is $1 - \epsilon_{\mathsf{pos}}$, where ϵ_{pos} is the false positive rate (see Definition 4).

Intuitively, a watermarking scheme $\Pi = (\mathsf{Watermark}, \mathsf{Detect})$ is considered robust and secure (for a meaningful choice of ϵ) if there does not exist a computationally efficient adversary A that satisfies the aforementioned Definition 5. The notion of "computationally efficient" depends on the context, but at the very least the adversary should be able to: (1) make black-box queries to the watermarked model itself, (2) perform efficient computations that do not require white-box access to the model or the data that it was trained on. As we demonstrate in Appendix D, there is a practical and universal adversary that meets Definition 5's criteria in several settings. However, if the adversary is restricted in other ways (for example only allowed to modify the output up to ϵ distance in some metric) then it may be possible to resist erasure attacks. This is the setting of "weak watermarking" that we mention in Section 1. We do not investigate weak watermarking schemes in this paper.

On quality approximation/degradation. According to Definition 5, an attack is considered successful when the quality of the non-watermarked output y' (computed by the adversary) is at least that of the target watermarked output $y \xleftarrow{r} \mathsf{M}_{\mathsf{k}}(x)$, i.e., $\mathsf{Q}(x, y') \geq \mathsf{Q}(x, y)$. We highlight that we can relax Definition 5 to consider adversaries that can remove the watermark at the price of some "small" degradation in the quality. This can be accomplished by modifying Definition 5 and requiring that $\mathsf{Q}(x, y') \geq \mathsf{Q}(x, y) - \gamma$ (instead of $\mathsf{Q}(x, y') \geq \mathsf{Q}(x, y)$) where γ is the parameter defining

[6] Here, we are implicitly assuming that the appended sentence does not significantly degrade the quality of the response.

the quality degradation tolerated by the definition. For simplicity, in this work we focus on the simpler setting in which there is no degradation, and the attacker has access to an exact quality oracle. In practice, the quality oracle may not be perfect, and hence some approximation and degradation might be necessary.

D IMPOSSIBILITY RESULT

We now show that a secret-key watermarking scheme can be defeated (i.e., remove the watermark) when the outputs of the underlying generative model can be perturbed. With the term "perturbed" we mean that the output can be slightly modified without affecting the quality. Examples of perturbable outputs are texts in which a word/token/span can be modified without changing the meaning of the text, or images in which a single pixel/patch can be modified (e.g., slightly change the RGB value of the target region) without affecting the picture. Thus, the impossibility result of this theorem applies to LLMs or image models.

D.1 ASSUMPTIONS AND DISCUSSION

Our impossibility result relies on the following two assumptions:

1. **Adversaries can check their own work:** The adversary has access to the quality oracle $\mathsf{Q} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ (associated to the target generative model $\mathsf{M} : \mathcal{X} \to \mathcal{Y}$).
2. **Random walk on the output space $\mathcal{Y}$:** At a high level, we assume that there is some graph $\mathsf{G} = (\mathcal{V}, \mathcal{E})$ in the output space $\mathcal{V} = \mathcal{Y}$ of the target watermarked model $\mathsf{M}_\mathsf{k} : \mathcal{X} \to \mathcal{Y}$. In addition, the adversary has access to an oracle P (dubbed *perturbation oracle*) that, on input $y \in \mathcal{Y}$ (e.g., the watermarked output), returns a random neighbor y' of y according to the graph G. For example, in the case of language, this can be obtained by taking some subset of the tokens and replacing them with random choices (either uniform or informed by large or smaller language models) or semantically equivalent spans. The main assumption we make is that the random walk has non-trivial *mixing* properties, and in particular that the graph is *irreducible* and *aperiodic* (these two conditions guarantee that a long-enough random walk will converge to its stationary distribution).

The definition of a perturbation oracle is given below.

Pertubation oracle. We abstract the ability of the adversary to perform a random walk over the output space of a (possibly watermarked) generative model $\mathsf{M} : \mathcal{X} \to \mathcal{Y}$ by introducing the notion of *perturbation oracles*. In a nutshell, a perturbation oracle $\mathsf{P} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ is a randomized oracle that, on input $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, returns a high-quality $y' \in \mathcal{Y}$ (according to some distribution) with probability at least ϵ_{pert}. The formal definition follows.

Definition 6 (Perturbation oracle)**.** A perturbation oracle P is a randomized oracle that, on input $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, returns an $y' \in \mathcal{Y}$. We say a perturbation oracle $\mathsf{P} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ with quality function $\mathsf{Q} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ is ϵ_{pert}*-preserving* if for every prompt $x \in \mathcal{X}$, for every $y \in \mathcal{Y}$, we have

$$\Pr[\mathsf{Q}(x, \mathsf{P}(x, y)) \geq \mathsf{Q}(x, y)] \geq \epsilon_{\mathsf{pert}}.$$

D.1.1 MIXING CONDITION

To define the mixing condition, we need to define an associated random walk for the perturbation oracle with respect to some particular prompt. We then consider the *mixing time* of this random walk (see Appendix G for a formal definition of mixing time). Fix a prompt $x \in \mathcal{X}$. The perturbation oracle $\mathsf{P}(x, \cdot)$ can be represented using a weighted directed graph $\mathsf{G}_x = (\mathcal{V}_x, \mathcal{E}_x)$ where an edge $(y_0, y_1) \in \mathcal{E}_x$ if weighted with the probability of going from output y_0 to output y_1 using $\mathsf{P}(x, \cdot)$. Below, we formally define two different graph (hierarchically ordered) representations corresponding to P. The first, named the$x$*-prompt graph representation*, is essentially the graph $\mathsf{G}_x = (\mathcal{V}_x, \mathcal{E}_x)$ described above, i.e., the graph representing the perturbation oracle $\mathsf{P}(x, \cdot)$ for a fixed prompt $x \in \mathcal{X}$. The second, named q*-quality* x*-prompt graph representation*, is the subgraph $\mathsf{G}_x^{\geq q} = (\mathcal{V}_x^{\geq q}, \mathcal{E}_x^{\geq q})$ of $\mathsf{G}_x = (\mathcal{V}_x, \mathcal{E}_x)$ where we only consider vertices of quality at least q, i.e., $\mathsf{Q}(x, y) \geq q$.

Definition 7 (Graph representation of perturbation oracles)**.** Let $\mathsf{P} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ be a perturbation oracle with associated quality function $\mathsf{Q} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$.

x-prompt graph representation: Fix a prompt $x \in \mathcal{X}$, the *x-prompt* graph representation of P is a weighted directed graph $\mathsf{G}_x = (\mathcal{V}_x, \mathcal{E}_x)$ such that $\mathcal{V}_x = \mathcal{Y}$ and $\mathcal{E}_x \subseteq \mathcal{Y} \times \mathcal{Y}$ defined as follows:

$$\mathcal{E}_x = \{(y_0, y_1) \in \mathcal{Y} \times \mathcal{Y} \ : \ \Pr[y_1 = \mathsf{P}(x, y_0)] > 0\} .$$

The weight function $\mathsf{weight}(y_0, y_1)$ of $\mathsf{G}_x = (\mathcal{V}_x, \mathcal{E}_x)$ is defined as follows:

$$\mathsf{weight}(y_0, y_1) = \begin{cases} \Pr[y_1 = \mathsf{P}(x, y_0)] & \text{if } (y_0, y_1) \in \mathcal{E}_x, \\ 0 & \text{otherwise.} \end{cases}$$

q-quality x-prompt graph representation: Fix $q \in [0, 1]$ and a prompt $x \in \mathcal{X}$, the *q-quality x-prompt* graph representation of P is a graph $\mathsf{G}_x^{\geq q} = (\mathcal{V}_x^{\geq q}, \mathcal{E}_x^{\geq q})$ such that $\mathcal{V}_x^{\geq q} \subseteq \mathcal{Y}$ and $\mathcal{E}_x^{\geq q} \subseteq \mathcal{Y} \times \mathcal{Y}$ defined as follows:

$$\mathcal{E}_x^{\geq q} = \{(y_0, y_1) \in \mathcal{Y} \times \mathcal{Y} \ : \ \mathsf{Q}(x, y_0) \geq q, \ \mathsf{Q}(x, y_1) \geq q, \text{and } \Pr[y_1 = \mathsf{P}(x, y_0)] > 0\} , \text{ and}$$

$$\mathcal{V}_x^{\geq q} = \{y \in \mathcal{Y} \ : \mathsf{Q}(x, y) \geq q\} .$$

The weight function $\mathsf{weight}(y_0, y_1)$ of $\mathsf{G}_x^{\geq q} = (\mathcal{V}_x^{\geq q}, \mathcal{E}_x^{\geq q})$ is defined as follows:

$$\mathsf{weight}(y_0, y_1) = \begin{cases} \Pr[y_1 = \mathsf{P}(x, y_0)] & \text{if } (y_0, y_1) \in \mathcal{E}_x^{\geq q}, \\ 0 & \text{otherwise.} \end{cases}$$

D.1.2 Discussion

The quality-oracle assumption is based on the intuition that it is easier, or at the very least not harder, to verify outputs than to generate them. In practice, we have used both zero-shot prompting (see Appendix E) and reward models to compute the quality function. In the context of images, it is also possible to use measures such as Inception score (Salimans et al., 2016), FID score (Heusel et al., 2017) or CLIP score (Hessel et al., 2021) etc. While at the moment zero-shot prompting works only for language models and not for audio/visual data, as models grow in power, both in capabilities and modalities, computing quality functions will only become easier.

The perturbation-oracle assumption is more subtle. First, note that the notion of "perturbation" can be abstract and does not need to correspond to changing a discrete subset of the output (e.g., a sequence of tokens or a patch of an image). The assumption can potentially fail for some prompts. For example, if a prompt has the form "Solve for the following set of n linear equations in n variables, and format the solution as ..." then there is zero entropy in the space of possible high-quality solutions. In such a case there is no "perturbation oracle" that can be defined. However, in such a case no watermarking is possible either. Indeed, watermarking schemes require some sort of "perturbation oracle" as well. To ensure successful watermarking, there needs to be a large set of potential high-quality solutions, and the scheme needs to be able to select a small subset of it (with a measure bounded by ϵ_{pos}— the false positive rate) based on the key. Hence we believe this assumption will be justified in cases where watermarking is feasible as well. Note that our perturbation oracle is in one crucial case much weaker than what is needed for watermarking: we do *not* require it to always preserve quality or even do so with high probability. It is enough that the probability of preserving quality is bounded away from zero.

Another potential issue with the perturbation oracle is that there could be outputs that are high-quality but would be very unlikely to be ever output from the model, whether watermarked or not. Hence the graph might be disconnected simply because of these "unreachable outputs". However, these can be addressed by introducing weights on the vertices as well that correspond to the probability of the output under the model. Such weighing makes no difference to our argument, but we omit this for clarity of notation.

D.2 Proof Overview of the Impossibility Result

The idea behind the impossibility result is to allow the adversary to perform a random walk on the graph while walking only over vertices with sufficiently high quality. Such a high-quality random

walk can be executed by leveraging the quality oracle Q and checking the quality of the next perturbed vertex (at each step of the random walk). Somewhat more formally, let x be a prompt and $\mathcal{V}_x^{\geq q}$ be a subset of vertices (of the graph) of quality $\mathsf{Q}(x, y) \geq q$. Assuming the vertices $\mathcal{V}_x^{\geq q}$ are *connected* (i.e., there is a non-zero probability of reaching any vertex in $\mathcal{V}_x^{\geq q}$), then an adversary, starting from a watermarked output $y \in \mathcal{Y}$ with quality $\mathsf{Q}(x, y) \geq q$, can perform a random walk (using its corresponding oracle) to reach a $y' \in \mathcal{V}_x^{\geq q}$ that is not watermarked. As we will see next, the existence of such a non-watermarked $y' \in \mathcal{V}_x^{\geq q}$ is guaranteed by the false positive rate of the watermarking scheme (see Definition 4).

Impossibility result. For simplicity, let us assume there exists a perturbation oracle P that is 1-preserving, i.e., for a fixed prompt $x \in \mathcal{X}$, the oracle *always* outputs a perturbed output $y' \in \mathcal{Y}$ whose quality is at least that of the initial value $y \in \mathcal{Y}$. Now, consider an adversary A with oracle access to P. The adversary's objective is to perturb the watermarked output y_0 t times (for some large enough t) to produce intermediate perturbations $y_1, \ldots, y_t$, such that the final perturbed output y_t will be independent of the watermarking secret key k. In this way, y_t will likely be verified as non-watermarked (i.e., $\mathsf{Detect}_{\mathsf{k}}(x, y_t) = 0$) with probability at least $1 - \epsilon_{\mathsf{pos}}$ where ϵ_{pos} is the false positive rate of the watermarking scheme. This can be achieved by requiring the adversary to compute $y_i \xleftarrow{r} \mathsf{P}(x, y_{i-1})$ for $i \in [t]$, where y_0 corresponds to the initial watermarked output. This adversarial strategy is effective in removing the watermark for the following two reasons:

- The interaction with the perturbation oracle can be represented as a t-step random walk over the q_0-quality x-prompt graph representation $\mathsf{G}_x^{\geq q_0} = (\mathcal{V}_x^{\geq q_0}, \mathcal{E}_x^{\geq q_0})$ of P where q_0 is the quality of the starting (watermarked) output y_0. Thus, $y_0, \ldots, y_t$ exactly represents the path taken by the adversary during such a random walk (note that each $y_1, \ldots, y_t$ will have the same quality q_0 by definition of a 1-preserving perturbation oracle. See Definition 6).
- If $\mathsf{G}_x^{\geq q_0}$ is irreducible and aperiodic, a random walk over $\mathsf{G}_x^{\geq q_0}$ converges to its unique stationary distribution $\vec{\pi}$ (see Appendix G.2).[7] In addition, if t corresponds to (or is slightly larger than) the ϵ_{dist}-mixing time of $\mathsf{G}_x^{\geq q_0}$, then the distribution of the final perturbed output y_t is ϵ_{dist}-close to the stationary distribution $\vec{\pi}$ of $\mathsf{G}_x^{\geq q_0}$. Furthermore, this distribution $\vec{\pi}$ is completely independent of the secret key k of the watermarking scheme since the secret key k is sampled (by the watermarking algorithm $\mathsf{Watermark}(\mathsf{M})$) independently from P's definition.

The above two facts, combined with the false positive ϵ_{pos}-rate of the watermarking scheme, directly imply that y_t will be classified as non-watermarked with probability at least $(1 - \epsilon_{\mathsf{pos}})(1 - \epsilon_{\mathsf{dist}})$. Moreover, we can set ϵ_{dist} to be arbitrarily small to achieve an adversarial advantage of $(1-\epsilon_{\mathsf{pos}})(1-\epsilon_{\mathsf{dist}}) \approx (1 - \epsilon_{\mathsf{pos}})$.[8]

The aforementioned approach relies on the fact that the perturbation oracle is 1-preserving; thus, each oracle invocation will certainly correspond to a new step of the random walk over the graph $\mathsf{G}_x^{\geq q_0} = (\mathcal{V}_x^{\geq q_0}, \mathcal{E}_x^{\geq q_0})$. This is because, starting from the watermarked output y_0 with quality q_0, each invocation will output (with probability 1) a perturbed output y_i with at least the same quality q_0. To make the impossibility more generic, we extend the attack to deal with perturbation oracles that may return low-quality perturbed values with *non-zero* probability, i.e., the perturbation oracle is ϵ_{pert}-preserving for some $\epsilon_{\mathsf{pert}} < 1$. In such a setting, we are not guaranteed that each invocation of P will correspond to a step of the random walk over $\mathsf{G}_x^{\geq q_0}$. Indeed, with probability $1 - \epsilon_{\mathsf{pert}}$, the ϵ_{pert}-preserving perturbation oracle may return $y' \in \mathcal{Y}$ such that $\mathsf{Q}(x, y') < q_0$, i.e., $y' \notin \mathcal{V}_x^{\geq q_0}$. To overcome this difficulty, we give the adversary oracle access to the quality function Q. In such a way, at each iteration $i \in [t]$, the adversary can check the quality of the perturbed value $y_i \xleftarrow{r} \mathsf{P}(x, y_{i-1})$ and proceed as follows:

- If $\mathsf{Q}(x, y_i) \geq q_0$, the adversary knows that the obtained perturbed value y_i corresponds to a new random walk step over $\mathsf{G}_x^{\geq q_0}$. Hence, in the next iteration, it will perturb the recently obtained output y_i (as described previously for the case of a 1-preserving perturbation oracle).

[7] Irreducibility and aperiodicity are the fundamental mixing properties that P (and its corresponding graph) must satisfy.

[8] Indeed, ϵ_{dist} approaches 0 exponentially as t increases (see Theorem 5).

- On the other hand, if $\mathsf{Q}(x, y_i) < q_0$, the value y_i *does not correspond* to a new random walk step. In this case, the adversary needs to re-send y_{i-1} to P until it obtains a perturbed value y_i such that $\mathsf{Q}(x, y_i) \geq q_0$.

Hence, by accessing the quality oracle Q, the adversary can continue walking over the high-quality graph $\mathsf{G}_x^{\geq q_0}$ even if the perturbation oracle is not errorless (i.e., $\epsilon_{\mathsf{pert}} < 1$). Naturally, when $\epsilon_{\mathsf{pert}} < 1$, the adversary needs to submit more queries to the perturbation oracle to overcome the low-quality outputs of P. This is taken into account by our theorem which relates the number of queries needed to the parameter ϵ_{pert}, i.e., the higher (resp. lower) ϵ_{pert}, the lower (resp. higher) the number of queries.

Theorem 2 reports our impossibility result that bounds A's advantage ϵ in erasing the watermark (for every $\mathsf{M} \in \mathcal{M}$ in a particular class $\mathcal{M} = \{\mathsf{M}_i : \mathcal{X} \to \mathcal{Y}\}$) when A has oracle access to a perturbation oracle P. As discussed in this section, such a result relies on the fact that P's graph representation is irreducible and aperiodic, which are necessary conditions to converge to its unique stationary distribution during the random walk. To make explicit the independence between the adversary (including its perturbation oracle P) and the secret key k of the target watermarked model M_{k}, we define P's irreducibility and aperiodicity properties w.r.t. the median quality among those of all possible responses (to a prompt x) that can be obtained by watermarking $\mathsf{M} \in \mathcal{M}$. Observe that the median quality does not depend on the secret key k used during the watermarking process but instead depends on *all possible choices* of secret keys (i.e., the key space of $\mathsf{Watermark}$). Formally, let $\mathcal{Q}_{\mathsf{M},x} = \{q_1, q_2, \ldots\}$ and q_{min} be defined as follows:

$$\mathcal{Q}_{\mathsf{M},x} = \left\{ q : \Pr\left[\mathsf{Q}(x, \mathsf{M}_{\mathsf{k}}(x)) = q : (\mathsf{k}, \mathsf{M}_{\mathsf{k}}) \xleftarrow{r} \mathsf{Watermark}(\mathsf{M}) \right] > 0 \right\}, \tag{1}$$

$$q_{\mathsf{min}} = \min_{\mathsf{M} \in \mathcal{M}, x \in \mathcal{X}} \{q_{\mathsf{M},x}\} \text{ where } q_{\mathsf{M},x} \text{ is the median quality of } \mathcal{Q}_{\mathsf{M},x}, \tag{2}$$

where the probability in Equation (1) is taken over the random coins of $\mathsf{Watermark}$ and M_{k}.[9] Then, the following Theorem 2 defines P's properties w.r.t. q_{min}, i.e., the minimum among all median qualities (see Equation (2)).

Theorem 2. *Let $\Pi = (\mathsf{Watermark}, \mathsf{Detect})$ be a watermarking scheme for a class of generative models $\mathcal{M} = \{\mathsf{M}_i : \mathcal{X} \to \mathcal{Y}\}$ with an associated quality function $\mathsf{Q} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$. Let $\mathsf{P} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ be a perturbation oracle (defined over the same prompt space $\mathcal{X}$ and output space $\mathcal{Y}$ as the class $\mathcal{M}$) with the same associated quality function $\mathsf{Q} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ as Π.*

Under the following conditions

1. *The watermarking scheme Π has a false positive ϵ_{pos}-rate;*
2. *The perturbation oracle P is ϵ_{pert}-preserving;*
3. *For every non-watermarked model $\mathsf{M} \in \mathcal{M}$, for every prompt $x \in \mathcal{X}$, for every quality $q \in [q_{\mathsf{min}}, 1]$, the q-quality x-prompt graph representation $\mathsf{G}_x^{\geq q}$ of P is irreducible and aperiodic where q_{min} is the minimum median defined in Equation (2). Also, let $t_{\mathsf{M},x}$ be the ϵ_{dist}-mixing time of $\mathsf{G}_x^{\geq q}$ where $\epsilon_{\mathsf{dist}} \approx 0$, and let t_{max} be the largest mixing time among $\{t_{\mathsf{M},x}\}_{\mathsf{M} \in \mathcal{M}, x \in \mathcal{X}}$;*

there exists an oracle-aided *universal adversary $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}$ that ϵ-breaks Π (Definition 5) by submitting at most $O(t_{\mathsf{max}}/\epsilon_{\mathsf{pert}})$ queries to P where $\epsilon \approx (1 - \epsilon_{\mathsf{pos}})/2$.*

The following corollary is obtained from Theorem 2 assuming (i) the perturbation oracle is $\frac{1}{2}$-preserving, and (ii) the watermarking scheme has a false positive rate of $\frac{1}{10}$.

Corollary 1. Let $\Pi = (\mathsf{Watermark}, \mathsf{Detect})$ be a watermarking scheme and P be a perturbation oracle as defined in Theorem 2. If Π has a false positive $\frac{1}{10}$-rate and P is $\frac{1}{2}$-preserving, then there exists an oracle-aided universal adversary $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}$ that breaks Π with a success probability of approximately $\frac{9}{20}$, by submitting at most $O(t_{\mathsf{max}})$ queries to P, where t_{max} is defined in Theorem 2.

[9] This is equivalent to saying that the quality values in $\mathcal{Q}_{\mathsf{M},x} = \{q_1, q_2, \ldots\}$ are defined over all possible choices of secret keys k (output by $\mathsf{Watermark}$) and all possible random coins of the resulting watermarked model M_{k}.

On improving A's advantage. In Theorem 2, A's advantage is approximately $\frac{1-\epsilon_{\mathsf{pos}}}{2}$ whereas $1-\epsilon_{\mathsf{pos}}$ is the maximum advantage that an adversary can achieve. The multiplicative $\frac{1}{2}$ loss is due to the definition of P w.r.t. to the minimum median q_{min}. Without loss of generality, it is possible to get an adversarial advantage ϵ arbitrarily close to $1-\epsilon_{\mathsf{pos}}$ by enforcing P's properties over a larger range of q values (Item 3 of Theorem 2). This can be accomplished by decreasing q_{min} (instead of using the median quality) such that convergence will be guaranteed for *almost* all watermarked outputs y (the ones with quality at least q_{min}).[10] For this reason, in Appendix H, we include Theorem 6 (and its corresponding proof) that is the extended version of the above theorem. In particular, Theorem 6 defines P's properties w.r.t. v-th quality percentile (instead of the median) so that, by decreasing v, we can get arbitrarily close to the best possible adversarial advantage $1-\epsilon_{\mathsf{pos}}$. Moreover, Theorem 6 explicates the concrete relation between the adversarial advantage ϵ, the number of queries, and the ϵ_{pert}-preservation of P. We refer the reader to Appendix H for more details.

E EXPERIMENTS

E.1 ATTACK IMPLEMENTATION

As a proof of concept empirical test of our attack schema, we construct an implementation of it and use it to attack several published watermarking schemes. While our general attack framework applies across all modalities, our focus is language, where quality oracles are currently more straightforward to implement. To implement our attack, we need to choose (1) the generative model and (2) the watermarking scheme, as well as implement (3) the perturbation oracle, and (4) the quality oracle.

We choose Llama2-7B (Touvron et al., 2023) as our generative model and attack three different watermarking schemes (Kirchenbauer et al., 2023a; Kuditipudi et al., 2023; Zhao et al., 2023a). We implement our *perturbation oracle* using T5-XL v1.1 (Raffel et al., 2020).[11] After trying out various potential reward model implementations, we settled on using an open-source reward model (RoBERTa-v3 large (Liu et al., 2019) fine-tuned on OpenAssistant (Köpf et al., 2023) preference data) as the primary quality oracle.[12]

This model occasionally misses simple errors in the text such as capitalization and punctuation mistakes and incoherent or irrelevant phrases, so we ask GPT-3.5 to perform a final check of the response quality with a focus on these sorts of errors, whenever the output is approved by the reward model (see Figure 8 for the prompt).[13] See Appendix J.1 for more details on the experimental setup.

All three of the watermarking schemes we attack were originally tested on the Real News subset of the C4 dataset (Raffel et al., 2020), so we use this as our primary task as well. Specifically, we give Llama-2-7B (Touvron et al., 2023) the task of generating a completion given the first 20 tokens of a news article. Except where otherwise noted, we default to a generation length of 200 tokens.

We test the following three watermark frameworks and refer readers to Appendix J.2 for more details:

- UMD (Kirchenbauer et al., 2023a) selects a randomized set of "green" tokens before a word is generated, and then softly promotes the use of green tokens during sampling, which can be detected efficiently.

[10] For example, by setting $q_{\mathsf{min}} = 0$, we obtain that the attack succeeds independently from the quality of the received initial watermarked output $y \xleftarrow{r} \mathsf{M}_{\mathsf{k}}(x)$. This is because the quality of y is always non-negative and, for every quality $q \in [0, 1]$, the q-quality x-prompt graph representation of P will be irreducible and aperiodic.

[11] This model is only pre-trained on C4 excluding any supervised training, to mask and infill one random span at each iteration. This design choice might be important as we want to generate a non-watermarked example without substantially shrinking the text length while the original T5 tends to infill texts with shorter lengths than the masked contents.

[12] https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2

[13] For each judgment, we calculate the scores of the original watermarked response and candidate response by passing the output of the reward model through a softmax layer, since this is how the reward model was supervised. We count the comparison as a tie if the difference in scores is less than $\Delta = 0.02$.

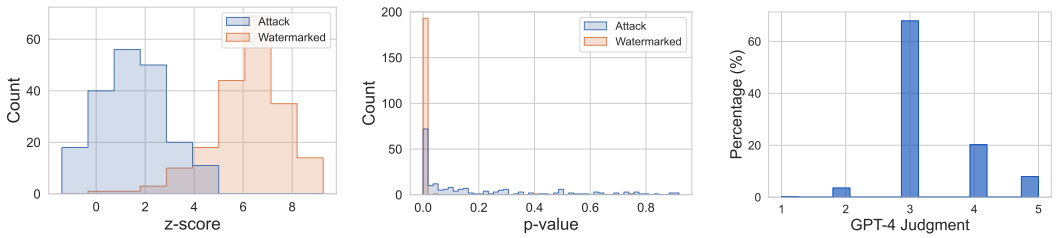
- EXP (Kuditipudi et al., 2023) is a distortion-free watermark framework that preserves the original LM's text distribution (over the randomness of the secret key), at least up to some maximum number of generated tokens.
- Unigram (Zhao et al., 2023a) is a watermarking scheme based on UMD designed to be provably robust to perturbations with bounded edit distance from the original watermarked text.

We also tested the watermarking schemes above on the Llama-2-7B-Chat model (Touvron et al., 2023). We observed that, likely due to the reduced entropy of this model in comparison to the base model, the schemes are often not successful at planting a watermark in the first place. For this reason, in the body of the paper we focus on the base model Llama-2-7B. In Appendix I.1 we show results for Llama-2-7B-Chat. We tested this model on long-form question-answering Fan et al. (2019) in addition to news article completion because it is capable of following instructions.

E.2 Experimental Results

We include our results on C4 news completion tasks with three watermark frameworks in Table 1. We show that with enough iterations, we can degrade watermark average detection performance to a z-score below 1.645 and a p-value greater than 0.05, a standard threshold that entails no more than 5% false positive rates. We also query GPT-4 to evaluate the comparative quality of the original watermarked output and the output after our attack. To do this, we prompt GPT-4 to compare the two outputs on a scale from 1 to 5 (corresponding to whether GPT4 "strongly" prefers one output, "slightly" prefers it, or judges them to "have similar quality", see Figure 9). Because GPT-4 has a bias towards preferring the first of the two outputs (Zheng et al., 2023), we query it twice, with both orderings of the two outputs.

We take a closer look at the UMD results by plotting histograms of the detection and GPT-4 judgment statistics for around 200 examples (Figure 3). These demonstrate that our attack can substantially degrade detection performance while typically inducing only minor quality degradation. Note that we choose the number of random walk steps for our attack to be high enough to push the average p-value above 0.05; we expect that running for longer would result in even higher p-values.



(a) Detection performance

(b) GPT-4 judgment

Figure 3: (a) Detection performance before (Watermarked) and after (Attack) our attack using Llama2-7B with UMD (Kirchenbauer et al., 2023a). (b) Comparative evaluation on watermarked texts against texts post-attack using GPT-4 as a judge. Scoring criteria: 1=post-attack response is much better than watermarked one, 2=slightly better, 3=of similar quality, 4=slightly worse, 5=much worse. Each example is included as two data points, one for each ordering of the two outputs in the GPT-4 query.

Detection performance and text quality over the course of the random walk. We also study how the watermark z-score and GPT-4 judge quality score change over time during the attack. For the sake of efficiency, we ran these tests for only 12 randomly selected prompts from the dataset. The results are averaged over these 12 examples, and the GPT-4 judge results are averaged over the two output orderings as well. In Figure 2, as the number of traversal steps increases, the z-score steadily decreases while the average GPT-4 judgment score fluctuates around its initial value of 0.

In Appendix J.3 of the appendix, we showcase how the text, detection statistics, and GPT-4 judge score change over the course of the random walk attack for a single example.

Controlling for perturbation oracle quality. If the perturbation oracle (T5-XL v1.1) were strong enough to produce texts on its own that are comparable in quality to the watermarked model M (Llama-2-7B), then our attack would be trivial. As a baseline and to calibrate our GPT-4 judge, we generate 40 samples by asking T5 for text completion iteratively: we feed the C4 news prefix for text completion and concatenate the generated contents back as input until the length can match the watermarked response. Then we ask GPT-4 to compare the watermarked responses against the T5-generated counterparts using the quality judge prompt mentioned above. The results are stark: the judge strongly prefers the watermarked Llama-2-7b output over the T5 output for every single example. In other words, when the watermarked response is presented first, the judge always says it is "much better" than the T5 response; and when the T5 response is presented first, the judge always declares it "much worse". This indicates that the quality oracle component of the attack was crucial to its success.

The impact of response length on attack performance. We plot the average z-score as a function of T denoted as the token length of the generated text (Figure 4). Note that a length greater than 400 is much larger than the default settings UMD was tested on, and we observe that the z-score keeps increasing for longer sequences, indicating detection performance degradation. On the other hand, our attack plateaus, showing that we can remove the watermark for long texts which can be essential for many practical tasks such as long-form reasoning (Nye et al., 2021), essay writing, etc. Results for short texts (less than 200 tokens) are included in Figure 7 of the appendix.

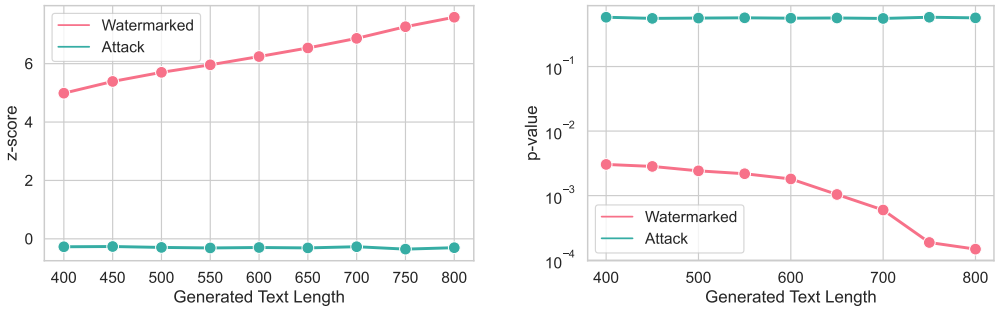


Figure 4: Detection performance and w.r.t. the watermarked text length using Llama2-7B-Chat with UMD (Kirchenbauer et al., 2023a). Results are aggregated across hundreds of examples.

E.3 Qualitative Examples

In Figure 5 we display two concrete non-cherrypicked before-and-after examples of the effects of our random walk attack on model outputs, so that the reader can get a sense of how quality is affected by the process. We use one prompt from LFQA Fan et al. (2019) and one SAT essay prompt. (For the SAT query, we use "*In any field of inquiry, the beginner is more likely than the expert to make important contributions. Write a response in which you discuss the extent to which you agree or disagree with the statement and explain your reasoning for the position you take. In developing and supporting your position, you should consider ways in which the statement might or might not hold true and explain how these considerations shape your position.*"). The comparison shows that the text after our attack can still be coherent, fluent, and on-topic. Such high-quality non-watermarked examples are abundant in our results. See Appendix I.3 for an example of the text at many intermediate rounds during the attack, with corresponding detection results and quality judge evaluations.

Tesla, Inc. is worth a lot of money for a number of reasons. Here are some of the key reasons: \n1. Disruptive Technology: Tesla's electric vehicles (EVs) are disrupting the auto industry, by providing a safer, sustainable, and greener transportation option. The company's innovative approach to EVs has captured the imagination of investors, while helping it secure future revenues. \n2. Growing Demand for EVs: As public awareness and regulation of vehicles and the environment continue to grow, there is an increasing demand for EVs. Tesla is well placed to benefit from this demand, thanks to its high-quality electric vehicles and growing charging network. \n3. Enhanced Financial Stability: Tesla's revenues come from various product categories, including vehicles, energy storage products, and services. The company's diversified revenue streams provide a more stable financial foundation than traditional automotive companies, which rely on vehicle sales alone. \n4. Strong Brand: Tesla has built a strong brand identity despite the challenges associated with electric vehicle manufacturing. The company's strong brand identity has helped it cultivate a growing customer base, while improving product pricing, manufacturing processes, and quality standards. \n5. Continuous Innovation: Tesla has a culture and process of continuous improvement, with new products and technologies announced regularly.

1 Tesla, Inc. is worth a lot of money due to a combination of factors. Here are some of the key reasons: \n1. Disruptive Technology: Tesla's electric vehicles (EVs) are disrupting the traditional automotive industry by offering a cleaner, more efficient, and technologically advanced alternative to gasoline-powered cars. The company's innovative approach to EVs has captured the imagination of consumers and investors alike. \n2. Growing Demand for EVs: As concerns about climate change and air pollution continue to grow, there is an increasing demand for EVs. Tesla is well-positioned to capitalize on this trend, given its strong brand, wide range of EV models, and growing charging network. \n3. Diversified Revenue Streams: Tesla generates revenue from several sources, including vehicle sales, energy products, and services. The company's diversified revenue streams provide a more stable financial foundation than traditional automotive companies, which rely primarily on vehicle sales. \n4. Strong Financials: Tesla has consistently delivered strong financial performance, with revenue and profitability growth in recent years. The company's commitment to cutting-edge technology and sustainability has attracted a loyal customer base, which has helped drive revenue and profitability. \n5. Continuous Innovation: Tesla's commitment to innovation has created a culture of continuous improvement, with new products and technologies in development.

In any field of inquiry, the beginner is king". This popularly believed scholarly statement sparks a heated scholarly debate. While some argue that beginner researchers are more creative and less burdened by existing knowledge, others imply that experts have had extensive experience facilitating them to make important and original contributions. In a way, it is both. It is undeniable that beginners often bring a fresh scholarly perspective. They challenge and transform the existing literature, delve deeper into existing concepts and formulate new ideas and theories. This creativity enables the emergence of important knowledge that most of the time often depends on novelty, changing perspectives and challenging existing theories. However, it is also important to recognize that expertise

1 In any field of inquiry, the beginner is more likely than the expert to make important contributions. This statement sparks a heated debate among scholars and researchers. While some argue that beginners are more creative and less burdened by existing knowledge, others contend that experts have a deeper understanding of the subject matter and can provide more informed insights. In my opinion, the statement holds some truth but is not wholly accurate. Firstly, it is undeniable that beginners often bring a fresh perspective to a field of inquiry. They are not encumbered by preconceived notions or biases and are more likely to challenge established theories. This creativity can lead to new ideas and innovations that experts might have overlooked. For instance, many groundbreaking scientific discoveries were made by young researchers who were not afraid to challenge established theories. However, it is so important to recognize that expertise

Figure 5: Un-watermarked example after our attack (red) and the corresponding original watermarked text (green) for the query "*Q: How come Tesla is worth so much money?*" and an SAT essay writing prompt. We do a word-by-word comparison and the differences between the two texts are highlighted.

F Concluding Remarks

This work provides general impossibility results for strong watermarking schemes. Although our specific attack is not very efficient, its primary advantage lies in its generality. Furthermore, while our implementation targets language generation models, the concepts of quality and perturbation oracles apply to all generative contexts, including audio-visual data. We posit, without formal proof, that verification is simpler than generation; thus, a general enhancement in capabilities and flexibility is likely to benefit the attacker (i.e., through improved quality and perturbation oracles) more than the defender (i.e., through better planting and detection algorithms). Hence we are not optimistic about the feasibility of strong watermarking schemes that would prevent a determined attacker from removing the watermark while preserving quality. Nonetheless, *weak* watermarking schemes do exist and can be useful, especially for safeguarding against unintentional leaks of AI-generated data (e.g., into training sets) or in scenarios with minimal risk that assume a low-effort attacker. Moreover, watermarking is just one tool in the arsenal of safety efforts for generative AI modeling. Cryptography can also be used to establish provenance of data. Ultimately, in many instances of disinformation, it is crucial to verify the data's source rather than to confirm whether AI generated it. Thus our goal is to set practical expectations regarding what watermarking schemes can achieve, thereby contributing to the safer use of AI models.

G Mathematical Background and Random Walks

G.1 Notation

We use the notation $[n] = \{1, 2, \ldots, n\}$. Capital bold-face letters (such as $\mathbf{X}$) are used to denote random variables, small letters (such as x) to denote concrete values, calligraphic letters (such as $\mathcal{X}$) to denote sets, and serif letters (such as A) to denote algorithms. For a string $x \in \{0, 1\}$, we let $|x|$ be its length; if $\mathcal{X}$ is a set, $|\mathcal{X}|$ represents the cardinality of $\mathcal{X}$.

We denote with $\vec{x} \in \mathcal{X}^n$ the column vector of length n with elements from $\mathcal{X}$. Similarly, we denote with $\vec{X} \in \mathcal{X}^{n \times m}$ the matrix with n rows, m columns, and elements from $\mathcal{X}$. We write $\vec{x}^\top$ (resp. $\vec{X}^\top$) to denote the transpose vector of $\vec{x}$ (resp. the transpose matrix of $\vec{X}$). Given a vector $\vec{x} \in \mathcal{X}^n$ (resp. a matrix $\vec{X} \in \mathcal{X}^{n \times m}$), we write $\vec{x}(i) \in \mathcal{X}$ (resp. $\vec{X}(i, j)$) to denote the i-th element of $\vec{x}$

(resp. the element located in the i-th row and j-th column of $\vec{X}$). We say a vector $\vec{x} \in \mathbb{R}^n$ is a (probability) distribution if $\sum_i \vec{x}(i) = 1$; if $\vec{x} \in \mathbb{R}^n$ is a distribution, we use the notation $\vec{x}$ and $\mathbf{x}$ interchangeably (where $\vec{x}(i) = \mathbf{x}(i)$ for every $i \in [n]$). When x is chosen uniformly from a set $\mathcal{X}$, we write $x \xleftarrow{r} \mathcal{X}$; if $\mathbf{x}$ is a distribution, we write $x \xleftarrow{r} \mathbf{x}$ to denote the act of sampling x according to the distribution $\mathbf{x}$. That is, $\Pr\left[x = i : i \xleftarrow{r} \mathbf{x}\right] = \mathbf{x}(i)$.

G.2 Graphs, Random Walks, and Mixing Time

Next, we focus on *weighted directed* graphs $\mathsf{G} = (\mathcal{V}, \mathcal{E})$ composed of n vertices $\mathcal{V} = [n]$. We denote by $\mathsf{weight}(i, j)$ the weight of the edge $(i, j) \in \mathcal{E}$ from $i \in \mathcal{V}$ to $j \in \mathcal{V}$. We assume that $\mathsf{weight}(i, j) = 0$ if and only if $(i, j) \notin \mathcal{E}$. Moreover, we define $\mathsf{weight}(i, \star) = \sum_j \mathsf{weight}(i, j)$, i.e., $\mathsf{weight}(i, \star)$ represents the sum of the edges' weights with source vertex i.

Random Walks. A t-step random walk over G is a probabilistic process such that, at each step, a neighbor is selected from the neighbors of the current vertex, according to the weight distribution. The transition matrix $\vec{P} \in \mathbb{R}^{n \times n}$ of a random walk over a weighted directed graph $\mathsf{G} = (\mathcal{V}, \mathcal{E})$ is defined as $\vec{P}(i, j) = \frac{\mathsf{weight}(i,j)}{\mathsf{weight}(i,\star)}$, i.e., $\vec{P}(i, j)$ represents the probability (based on the graph's weights) of reaching vertex j from vertex i. Let $\mathbf{p}_0 \in \mathbb{R}^n$ be the starting distribution (of a random walk), and let $\mathbf{p}_t \in \mathbb{R}^n$ represent the distribution after t steps of the random walk. By definition, we have $\mathbf{p}_{t+1}(i) = \sum_{j:(j,i)\in\mathcal{E}} \mathbf{p}_t(j) \cdot \vec{P}(j, i)$, where $\mathbf{p}(i)$ is the probability associated with $i \in \mathcal{V}$. Similarly, the (global) distribution after a t-step random walk can be expressed as $\mathbf{p}_{t+1} = \vec{P}^\top \cdot \mathbf{p}_t$.

Below, we recall the definition of the stationary distribution of a random walk.

Definition 8 (Stationary distribution of random walks)**.** Let $\mathsf{G} = (\mathcal{V}, \mathcal{E})$ be a weighted directed graph and $\vec{P}$ be the transition matrix of G. We say that $\vec{\pi} \in \mathbb{R}^n$ is a *stationary distribution* for $\vec{P}$ if $\vec{P}^\top \cdot \vec{\pi} = \vec{\pi}$.

In other words, the above definition states that the probability distribution after a 1-step random walk is the stationary distribution $\vec{\pi}$ if the starting distribution $\mathbf{p}_0$ is the stationary distribution itself. By induction, this also implies that the distribution $\mathbf{p}_t$ after a t-step random walk is equal to the stationary distribution $\vec{\pi}$ if $\mathbf{p}_0 = \vec{\pi}$, for every $t \geq 1$.

The following theorem provides the stationary distribution for any weighted directed graph.

Theorem 3. *Let $\mathsf{G} = (\mathcal{V}, \mathcal{E})$ be a weighted directed graph and $\vec{P}$ be its transition matrix. The distribution $\vec{\pi} \in \mathbb{R}^n$ such that*

$$\forall i \in \mathcal{V},\ \vec{\pi}(i) = \frac{\mathsf{weight}(i, \star)}{\sum_{i\in\mathcal{V}} \mathsf{weight}(i, \star)} = \frac{\sum_{j\in\mathcal{V}} \mathsf{weight}(i, j)}{\sum_{i\in\mathcal{V}} \sum_{j\in\mathcal{V}} \mathsf{weight}(i, j)} = \frac{\sum_{j\in\mathcal{V}} \mathsf{weight}(i, j)}{\sum_{(i,j)\in\mathcal{E}} \mathsf{weight}(i, j)}$$

is a stationary distribution for $\vec{P}$.

Interestingly, the stationary distribution is unique when the underlying directed graph G is irreducible (i.e., the graph does not have leaf nodes). Moreover, if G is also aperiodic then, independently of the initial starting distribution $\mathbf{p}_0$, a random walk converges to its stationary distribution $\vec{\pi}$ as $t \to \infty$.

Theorem 4 (Convergence to the stationary distribution)**.** *If a weighted directed graph $\mathsf{G} = (\mathcal{V}, \mathcal{E})$ is irreducible and aperiodic, there exists a unique stationary distribution $\vec{\pi}$. Moreover, for every $\mathbf{p}_0 \in \mathbb{R}^n$, $\mathbf{p}_t = (\vec{P}^\top)^t \cdot \mathbf{p}_0$ (i.e., the probability distribution after a t-step random walk with starting distribution $\mathbf{p}_0$) converges to $\vec{\pi}$ as $t \to \infty$.*

In this work, we are interested in setting a bound for t in the above theorem. In more detail, we want to bound the minimum number of steps required by the random walk to get close enough to its stationary distribution. This is known as the ϵ_{dist}-mixing time, and it can be bounded using the *second largest eigenvalue* (in absolute value) of the transition matrix of the graph (note that the second largest eigenvalue also has connections with the conductance of the graph). Below, we report the formal definition of mixing time and its corresponding bound.

Definition 9 (ϵ_{dist}-mixing time)**.** Let $\mathsf{G} = (\mathcal{V}, \mathcal{E})$ be a weighted directed graph that is irreducible and aperiodic (as in Theorem 4) and let $\vec{P}$ be its corresponding transition matrix. For any $0 < \epsilon_{\mathsf{dist}} \leq 1$, the ϵ_{dist}*-mixing time* $t_{\mathsf{min}}(\epsilon_{\mathsf{dist}})$ of $\vec{P}$ is the smallest number of steps t such that for every starting distribution $\mathbf{p}_0 \in \mathbb{R}^n$, we have

$$\left|\mathbf{p}_t - \vec{\pi}\right| = \left|(\vec{P}^\top)^t \cdot \mathbf{p}_0 - \vec{\pi}\right| \leq \epsilon_{\mathsf{dist}},$$

where $\vec{\pi}$ is the unique stationary distribution of $\vec{P}$.

Theorem 5 (Bound on ϵ_{dist}-mixing time)**.** *Let $\mathsf{G} = (\mathcal{V}, \mathcal{E})$ be a weighted directed graph that is irreducible and aperiodic, and let $\vec{P}$ be its corresponding transition matrix. Also, let $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_n$ be the eigenvalues of $\vec{P}$, and let $\vec{\pi}$ be the unique stationary distribution of $\vec{P}$ (Theorem 4). For $g = \max\{|\alpha_2|, \ldots, |\alpha_n|\}$ and $\pi_{\min} = \min\{\vec{\pi}(1), \ldots, \vec{\pi}(n)\}$, the ϵ_{dist}-mixing time of $\vec{P}$ is*

$$t_{\min}(\epsilon_{\mathsf{dist}}) \leq O\left(\frac{1}{1-g} \cdot \log\left(\frac{1}{\pi_{\min} \cdot \epsilon_{\mathsf{dist}}}\right)\right).$$

H Extended Impossibility Theorem

In this section, we include an extended version of the impossibility result which explicates the concrete relation between the adversarial advantage ϵ, the number of queries, and the ϵ_{pert}-preservation of P. In addition, to make the result as generic as possible, we define P's properties (i.e., irreducibility and aperiodicity) w.r.t. any arbitrary v-th quality percentile (for $v \in [0, 100]$) of the quality values of all possible responses (of prompt x) that can be obtained by watermarking $\mathsf{M} \in \mathcal{M}$ (recall that in Theorem 2 we used the median instead of the v-th quality percentile).

Formally, for $v \in [0, 100]$, let $\mathcal{Q}_{\mathsf{M},x} = \{q_1, q_2, \ldots\}$ and q_{min} be defined as follows:

$$\mathcal{Q}_{\mathsf{M},x} \text{ is defined as in Equation (1),} \tag{3}$$

$$q_{\mathsf{min}} = \min_{\mathsf{M} \in \mathcal{M}, x \in \mathcal{X}} \{q_{\mathsf{M},x}\} \text{ where } q_{\mathsf{M},x} \text{ is the } v\text{-th quality percentile of } \mathcal{Q}_{\mathsf{M},x}. \tag{4}$$

Below, we report the extended version of the impossibility result whose perturbation oracle is defined w.r.t. q_{min} of Equation (4). The proof is given in Appendix H.1.

Theorem 6. *Let $\Pi = (\mathsf{Watermark}, \mathsf{Detect})$ be a watermarking scheme for a class of generative models $\mathcal{M} = \{\mathsf{M}_i : \mathcal{X} \to \mathcal{Y}\}$ with an associated quality function $\mathsf{Q} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$. Let $\mathsf{P} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ be a perturbation oracle (defined over the same prompt space $\mathcal{X}$ and output space $\mathcal{Y}$ of the class $\mathcal{M}$) with the same associated quality function $\mathsf{Q} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ as Π.*

Under the following conditions

1. *The watermarking scheme Π has a false positive ϵ_{pos}-rate (Definition 4);*
2. *The perturbation oracle P is ϵ_{pert}-preserving (Definition 6);*
3. *For every non-watermarked model $\mathsf{M} \in \mathcal{M}$, for every prompt $x \in \mathcal{X}$, for every quality $q \in [q_{\mathsf{min}}, 1]$, the q-quality x-prompt graph representation $\mathsf{G}_x^{\geq q}$ of P is irreducible and aperiodic where q_{min} is the minimum median defined in Equation (4) (for some arbitrary $v \in [0, 100]$). Also, let $\vec{\pi}_{x,q}$ be the unique stationary distribution[14] of the transition matrix $\vec{P}_{x,q}$ of $\mathsf{G}_x^{\geq q}$ (Definition 8) and, for $\epsilon_{\mathsf{dist}} \in [0, 1]$, let $t_{x,q}$ be the ϵ_{dist}-mixing time of $\vec{P}_{x,q}$ (Definition 9 and Theorem 5) defined as follows:*

$$t_{x,q} = \omega\left(\frac{1}{1 - \max\{|\alpha_2^{(x,q)}|, \ldots, |\alpha_n^{(x,q)}|\}} \cdot \log\left(\frac{1}{\pi_{\min}^{(x,q)} \cdot \epsilon_{\mathsf{dist}}}\right)\right)$$

where $\pi_{\min}^{(x,q)} = \min\{\vec{\pi}_{x,q}(1), \ldots, \vec{\pi}_{x,q}(n_{x,q})\}$, $n_{x,q} = |\mathcal{V}_x^{\geq q}|$, and $\alpha_1^{(x,q)} \geq \alpha_2^{(x,q)} \geq \ldots \geq \alpha_n^{(x,q)}$ are the eigenvalues of the transition matrix $\vec{P}_{x,q}$.[15]

[14] Recall that a random walk converges to its unique stationary distribution when the corresponding weighted directed graph is irreducible and aperiodic (Theorem 4).

[15] Observe that $t_{x,q}$ is asymptotically larger than the ϵ_{dist}-mixing time of $\mathsf{G}_x^{\geq q}$ as defined in Theorem 5.

there exists an oracle-aided *universal adversary* $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}$ *that* ϵ*-breaks* Π *(Definition 5) by submitting at most* t *queries to* P *where*

$$\epsilon = \left(1 - \frac{v}{100}\right)(1 - \epsilon_{\mathsf{pos}})(1 - \epsilon_{\mathsf{dist}})\left(1 - \sum_{k=0}^{t-t_{\mathsf{err}}-1} \binom{t}{k} (\epsilon_{\mathsf{pert}})^k (1 - \epsilon_{\mathsf{pert}})^{t-k}\right), \text{ and}$$
$$t = \max_{x \in \mathcal{X}, q \in [q_{\mathsf{min}}, 1]} \{t_{x,q}\} + t_{\mathsf{err}}.$$

Observe that the above theorem implies Theorem 2 since q_{min} is the minimum median quality (as in Theorem 2) when $v = 50$.

For sufficiently small values of v, ϵ_{dist}, and sufficiently high values of ϵ_{pert}, the adversarial advantage ϵ of Theorem 6 approaches $1 - \epsilon_{\mathsf{pos}}$ (i.e., the best possible advantage).[16] In turn, if ϵ_{pos} is also small (e.g., $\epsilon_{\mathsf{pos}} \leq \frac{1}{10}$), ϵ is close to 1.

H.1 Proof of Theorem 6

Fix P and Q as defined in Theorem 6. Let $\mathsf{M} \in \mathcal{M}$ and let $(\mathsf{k}, \mathsf{M}_{\mathsf{k}})$ (output by $\mathsf{Watermark}(\mathsf{M})$) be the original generative model and its watermarked version (with the associated secret-key k) that the adversary is trying to break (recall that M and $(\mathsf{k}, \mathsf{M}_{\mathsf{k}})$ are unknown to A). Consider the following adversary $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}$:

1. Receive a prompt $x \in \mathcal{X}$ and a watermarked output $y \in \mathcal{Y}$ (recall that y is computed as $y \xleftarrow{r} \mathsf{M}_{\mathsf{k}}(x)$ as defined in the erasure attack experiment of Definition 5).
2. Initialize $y_0 = y$ and $\mathsf{ctr} = 0$.
3. Send (x, y_0) to Q and receive the answer q_0.
4. For each $i \in [t]$, the adversary proceeds as follows (where t is as defined in Theorem 6):
   (a) Send (x, y_{i-1}) to P and receive the answer $\tilde{y}$.
   (b) Send $(x, \tilde{y})$ to Q and receive the answer $\tilde{q}$.
   (c) If $\tilde{q} \geq q_0$, set $y_i = \tilde{y}$ and increment the counter ctr (i.e., $\mathsf{ctr} = \mathsf{ctr} + 1$). Otherwise, if $\tilde{q} < q_0$, set $y_i = y_{i-1}$.
5. Finally, output y_t if $\mathsf{ctr} \geq t - t_{\mathsf{err}}$ (where t and t_{err} are as defined in Theorem 6). Otherwise, if $\mathsf{ctr} < t - t_{\mathsf{err}}$, output $\perp$ (i.e., an error message).

To show that $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}$ ϵ-breaks the watermarking scheme Π, we prove the following three lemmas.

Lemma 1. For every $\mathsf{M} \in \mathcal{M}$, for every prompt $x \in \mathcal{X}$, we have

$$\Pr[\mathsf{Q}(x, y_0) \geq q_{\mathsf{min}}] \geq 1 - \frac{v}{100},$$

where $y_0 = y$ is the watermarked output given as input to the adversary $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}$.

Proof. The lemma follows by observing that $q_{\mathsf{min}} = \min_{\mathsf{M} \in \mathcal{M}, x \in \mathcal{X}} \{q_{\mathsf{M},x}\}$ where $q_{\mathsf{M},x}$ is the v-th quality percentile of the list $\mathcal{Q}_{\mathsf{M},x}$ as defined in Equations (3) and (4). In other words, q_{min} is the minimum among the v-th percentiles $\{q_{\mathsf{M},x}\}$ each calculated over all possible random coins of both $\mathsf{Watermark}$ and the watermarked version of M. By definition, this implies that

$$\Pr[\mathsf{Q}(x, \mathsf{M}(x)) \geq q_{\mathsf{min}}] \geq 1 - \frac{v}{100}.$$

This concludes the proof of Lemma 1. □

[16] If ϵ_{pert} is not high enough, the same result can be achieved by increasing t_{err} of Theorem 6. Note that by increasing t_{err}, we increase the overall number t of perturbation queries.

Lemma 2. For every $\mathsf{M} \in \mathcal{M}$, for every prompt $x \in \mathcal{X}$, we have that

$$\Pr\left[\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}(x,y) \neq \perp\right] = 1 - \sum_{k=0}^{t-t_{\mathsf{err}}-1} \binom{t}{k} (\epsilon_{\mathsf{pert}})^k (1-\epsilon_{\mathsf{pert}})^{t-k},$$

where $y_0 = y$ is the watermarked output given as input to the adversary $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}$.

Proof. Note that $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}(x,y)$ outputs $\perp$ only if $\mathsf{ctr} < t - t_{\mathsf{err}}$. Moreover, *the counter* ctr *is not incremented* only when the perturbation oracle, on input (x, y_i) (for some $i \in [t]$), returns $\tilde{y}$ such that $\mathsf{Q}(x,\tilde{y}) < \mathsf{Q}(x,y_i)$. The latter occurs with probability at most $1-\epsilon_{\mathsf{pert}}$ since the perturbation oracle P is ϵ_{pert}-preserving (Definition 6).

Let $\mathbf{X}$ be the random variable describing the value of ctr at the end of the adversary's computation. Then, we have that

$$\Pr\left[\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}(x,y) \neq \perp\right] = \Pr[\mathbf{X} \geq t - t_{\mathsf{err}}] = 1 - \Pr[\mathbf{X} < t - t_{\mathsf{err}}] = 1 - \Pr[\mathbf{X} \leq t - t_{\mathsf{err}} - 1]. \tag{5}$$

The probability $\Pr[\mathbf{X} \leq t - t_{\mathsf{err}} - 1]$ is characterized by a binomial distribution where the probability of incrementing $\mathbf{X}$ (resp. not incrementing $\mathbf{X}$) is ϵ_{pert} (resp. $1-\epsilon_{\mathsf{pert}}$). Formally,

$$\Pr[\mathbf{X} \leq t - t_{\mathsf{err}} - 1] = \sum_{k=0}^{t-t_{\mathsf{err}}-1} \binom{t}{k} (\epsilon_{\mathsf{pert}})^k (1-\epsilon_{\mathsf{pert}})^{t-k}. \tag{6}$$

Lemma 2 follows by combining Equations (5) and (6). □

Lemma 3. For every $\mathsf{M} \in \mathcal{M}$, for every prompt $x \in \mathcal{X}$, conditioned on $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}(x,y) \neq \perp$ and $\mathsf{Q}(x,y) = q_0 \geq q_{\mathsf{min}}$, we have that

$$\left|\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}(x,y) - \vec{\pi}_{x,q_0}\right| \leq \epsilon_{\mathsf{dist}}$$

where $\vec{\pi}_{x,q_0}$ is the unique stationary distribution of the transition matrix $\vec{P}_{x,q_0}$ of $\mathsf{G}_x^{\geq q_0}$.

Proof. Assume that $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}(x,y) \neq \perp$ and $\mathsf{Q}(x,y) = q_0 \geq q_{\mathsf{min}}$. It is easy to see that the computation of the adversary $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}(x,y)$ is exactly a random walk over $\mathsf{G}_x^{\geq q_0} = (\mathcal{V}_x^{\geq q_0}, \mathcal{E}_x^{\geq q_0})$, where q_0 is the quality of the watermarked output $y = y_0$ given as input to the adversary. This is because at each iteration $i \in [t]$, the adversary sets $y_i = \tilde{y}$ (i.e., it moves from y_{i-1} to $y_i = \tilde{y}$ according to $\mathsf{G}_x^{\geq q_0}$) only if $\tilde{y}$ has a quality of at least q_0. This corresponds exactly to a random walk over the vertices with quality at least q_0, which is the definition of the q_0-quality x-prompt graph representation $\mathsf{G}_x^{\geq q_0}$ of P.

In addition, the following conditions hold:

1. By leveraging Item 3 of Theorem 6, for every $q \in [q_{\mathsf{min}}, 1]$, the q-quality x-prompt graph representation $\mathsf{G}_x^{\geq q}$ is irreducible and aperiodic. Thus, a random walk over the weighted directed graph $\mathsf{G}_x^{\geq q_0}$ will *eventually* converge to its unique stationary distribution $\vec{\pi}_{x,q_0}$ (recall that $q_0 \geq q_{\mathsf{min}}$ by assumption).

2. By assumption $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}(x,y) \neq \perp$. Thus, $\mathsf{ctr} \geq t - t_{\mathsf{err}} = \max_{x \in \mathcal{X}, q \in [q_{\mathsf{min}},1]}\{t_{x,q}\}$ (as defined in Item 3 of Theorem 6) which, in turn, implies $\mathsf{ctr} \geq t_{x,q_0}$ since $q_0 \geq q_{\mathsf{min}}$. Note that ctr corresponds to the number of steps performed by the adversary during the random walk over the q_0-quality x-prompt graph $\mathsf{G}_x^{\geq q_0} = (\mathcal{V}_x^{\geq q_0}, \mathcal{E}_x^{\geq q_0})$.

By leveraging the above two conditions, we conclude that (i) a random walk over $\mathsf{G}_x^{\geq q_0} = (\mathcal{V}_x^{\geq q_0}, \mathcal{E}_x^{\geq q_0})$ converges to its unique stationary distribution $\vec{\pi}_{x,q_0}$, and (ii) $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}(x,y)$'s random walk is composed of at least $\mathsf{ctr} \geq t_{x,q_0}$ steps where t_{x,q_0} is asymptotically larger than the ϵ_{dist}-mixing time of the transition matrix $\vec{P}_{x,q_0}$ of $\mathsf{G}_x^{\geq q_0}$ (see Item 3 of Theorem 6). Thus, we conclude that the output distribution of $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}(x,y)$ is ϵ_{dist}-close to $\vec{\pi}_{x,q_0}$, i.e.,

$$\left|\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}(x,y) - \vec{\pi}_{x,q_0}\right| \leq \epsilon_{\mathsf{dist}}.$$

This concludes the proof. □

By leveraging Lemmas 1 to 3, we have that for every $\mathsf{M} \in \mathcal{M}$, for every prompt $x \in \mathcal{X}$, the following conditions hold:

1. Let $\mathbf{E}$ be the event that $\mathsf{Q}(x, y) = q_0 \geq q_{\mathsf{min}}$ and $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}(x, y) \neq \perp$. Then, by leveraging Lemmas 1 and 2 we conclude that $\mathbf{E}$ occurs with probability at least

$$\left(1 - \frac{v}{100}\right)\left(1 - \sum_{k=0}^{t-t_{\mathsf{err}}-1} \binom{t}{k} (\epsilon_{\mathsf{pert}})^k (1 - \epsilon_{\mathsf{pert}})^{t-k}\right),$$

   where the probability is taken over (k, M) output by $\mathsf{Watermark}(\mathsf{M})$, the random coins of the watermarked model M_k, and the perturbation oracle P.

2. Conditioned on $\mathbf{E}$, the quality $\mathsf{Q}(x, y_t)$ of $y_t \neq \perp$ (output by the adversary) is at least $q_0 \geq q_{\mathsf{min}}$. This is because y_t is the result of a random walk over the graph $\mathsf{G}_x^{\geq q_0} = (\mathcal{V}_x^{\geq q_0}, \mathcal{E}_x^{\geq q_0})$ (of the perturbation oracle P) composed of all vertices of quality at least q_0 (see also the proof of Lemma 3).

3. Conditioned on $\mathbf{E}$, the output $y_t \neq \perp$ (produced by the adversary) is such that $\mathsf{Detect}_\mathsf{k}(x, y_t) = 0$ (i.e., y_t is not watermarked) with probability at least $(1-\epsilon_{\mathsf{pos}})(1-\epsilon_{\mathsf{dist}})$. This follows by observing that, conditioned on $\mathbf{E}$ (i.e., $\mathsf{Q}(x, y) = q_0 \geq q_{\mathsf{min}}$ and $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}(x, y) \neq \perp$), the output distribution of $\mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}(x, y)$ is ϵ_{dist}-close to the unique stationary distribution $\vec{\pi}_{x,q_0}$ (as defined in Lemma 3). In turn, since $\vec{\pi}_{x,q_0}$ is independent of $(\mathsf{k}, \mathsf{M}_\mathsf{k}) \xleftarrow{r} \mathsf{Watermark}(\mathsf{M})$,[17] we have that $\mathsf{Detect}_\mathsf{k}(x, y_t) = 0$ with probability at least $(1 - \epsilon_{\mathsf{pos}})$ due to the false positive ϵ_{pos}-rate (Definition 4) of the watermarking scheme Π.

By combining the above arguments, we conclude that, for every model $\mathsf{M} \in \mathcal{M}$, for every prompt $x \in \mathcal{X}$, the following probability holds:

$$\Pr\left[\ \mathsf{Detect}_\mathsf{k}(x, y_t) = 0 \text{ and } \mathsf{Q}(x, y_t) \geq \mathsf{Q}(x, y) \ : \ y \xleftarrow{r} \mathsf{M}_\mathsf{k}(x),\ y_t \xleftarrow{r} \mathsf{A}^{\mathsf{P}(\cdot,\cdot),\mathsf{Q}(\cdot,\cdot)}(x, y)\ \right] \geq \epsilon$$

where $(\mathsf{k}, \mathsf{M}_\mathsf{k}) \xleftarrow{r} \mathsf{Watermark}(\mathsf{M})$ and ϵ as defined in Theorem 6. This concludes the proof of Theorem 6.

I Additional Experimental Results

I.1 Additional Results with Chat Model

We include our results on two watermark schemes and two datasets using Llama-2-Chat (Table 2). In part due to the reduced entropies, the watermarks we selected did not work with the chat variant of Llama-2 even before our attack. Despite that, our attack can get the average degradation of a z-score of 1.492 and a p-value of 0.156. To be realistic, we set the maximum generated tokens of the watermarked texts as 512, a larger one reported in previous works. Notably, only UMD is robust to such a setting, but our attack can substantially reduce the average z-score down to the detection threshold 4, which evades the detectors for UMD (Kirchenbauer et al., 2023a) as shown in most samples in Figure 6. For Unigram, we can still reduce the z-score to a lower level.

Table 2: Results of our attack on three watermark schemes using Llama-2-Chat. Both results averaged across 512 examples before and after our attack are shown.

Model	Framework	C4 Real News		LFQA	
		z-score	p-value	z-score	p-value
Llama2-7B-Chat	UMD (Kirchenbauer et al., 2023a)	5.023 → 1.875	0.010 → 0.179	1.131 → -0.050	0.258 → 0.522
	Unigram (Zhao et al., 2023a)	2.391 → 1.554	0.118 → 0.192	2.038 → 1.236	0.180 → 0.298

[17] This is because the output distribution of the perturbation oracle P (which, in turn, defines its corresponding graph) is defined before the sampling of $(\mathsf{k}, \mathsf{M}_\mathsf{k})$ according to $\mathsf{Watermark}(\mathsf{M})$.

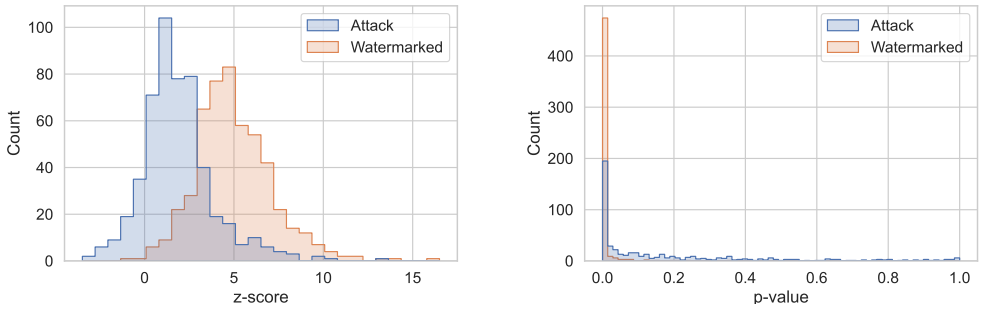


Figure 6: Detection performance before (Watermarked) and after (Attack) our attack using Llama2-7B-Chat with UMD (Kirchenbauer et al., 2023a). Results are based on 512 examples.

I.2 Results on Short Texts

As shorter responses might be easier for detection and automated evaluation, we change the length of the watermarked response to see how detection and attack performance would vary. We show results on texts with less than 200 tokens in Figure 7. As the generated text length increases, both text with watermark and after our attack would become harder to detect, such gap becomes larger as sequence length increases.

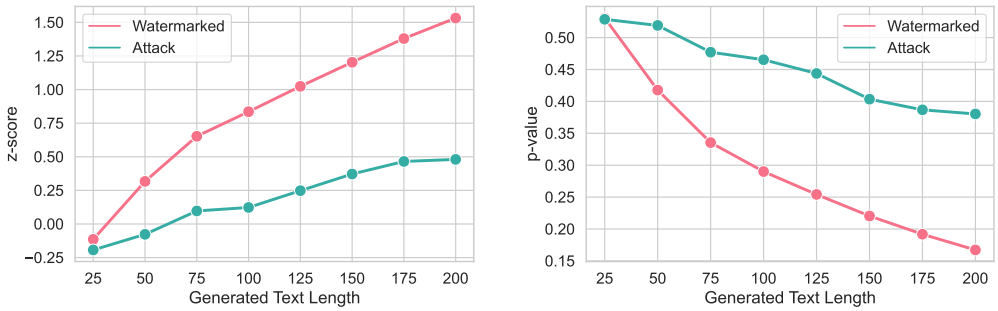


Figure 7: Detection performance with respect to watermarked sequence length using Llama2-7B-Chat with UMD (Kirchenbauer et al., 2023a).

I.3 Qualitative Results through Steps

Given a C4 news article prefix "*"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*", we showcase in Appendix J.3 the results of removing the watermark in the following response: *"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."*

Note that for the initial watermarked response, the detection algorithm of UMD gives a z-score of 7.340 and a p-value of 0.000. We report the results every 5 valid random walks. At each generation,

we highlighted the different parts of the texts that are modified with the corresponding detection (z-score, p-value) and GPT-4 quality judgment results.

We observe that the detection performance generally keeps decreasing while the new texts are of high quality according to the quality oracle implemented as GPT-4. Especially, the oracle score stays at 0 for the final several examples, showing that they are of similar quality to the watermarked response.

J EXPERIMENTAL DETAILS

J.1 IMPLEMENTATION DETAILS OF ATTACK

We discuss our key design choices and implementations. In general, we found our attack effective for all the settings considered and is not susceptible to the hyper-parameters and choices below.

Perturbation oracle. Recall that we generate watermarked texts with maximum generation tokens of 200 or 512 and attack by replacing one span of the text at a time, thus we set the span length to be 6 for all the attacks in the main table. For each infill, we do top-p sampling and p, the minimum. and maximum infilled text length according to Table 4. Note that we generally use the default hyper-parameters and don't tune them too much. We incorporate backtracking into the random walk as another error-reduction mechanism. If the perturbation oracle suggests many candidates (above some "patience" threshold) without any of them passing the quality checks, then we undo the most recent step of the walk.

Quality oracle. Note that in general all kinds of watermarks are removable by omitting contents but would degrade the quality. Therefore, it is important to ensure the quality at each step of our iterative attack process. To realize this goal, we implement three alternative quality oracles, trading off quality, cost, and efficiency (Table 3). In principle, attackers can tailor quality oracles according to their needs, considering the trade-offs among quality, efficiency, and cost. For example, i) Efficiency and Cost: malicious users can efficiently generate many high-quality texts using GPT-4 for automated phishing and strip the watermark using reward models as the quality oracle with slight degradation in quality. ii) Quality: a student can wait a week to generate a solution with GPT-4 to one assignment problem whose deadline is one week from now with hundreds of dollars for paying for GPT-4 as a quality oracle.

Though our impossibility results are generic and the assumptions can be made stronger when models become more capable, we instantiate our quality oracle reward models + GPT-3.5 for quality checking for most experiments. The reasons are twofold: firstly, it's much more efficient and less expensive to use reward models for comparing responses to filter obviously bad texts; moreover, we find that GPT-3.5 and GPT-4 have significant position biases (Zheng et al., 2023) that grade the first response with higher quality when evaluating two responses to a given query (prompting details are in Appendix J.3). We find that such position bias limitation is substantial in practice when the response length is greater than 200 so it'd be hard to get non-watermarked examples in a reasonable amount of queries even if powerful models can evaluate multiple nuance aspects of individual responses. We detail the trade-offs in Table 3.

Design choices to ensure text quality. When masking a span, we split and mask the words rather than tokens to avoid generating nonsensical words that degrade text quality. To alleviate the impact of position bias on the quality oracle, we query it twice and categorize the results as win, tie, or lose. Our goal is to get a new text we non-degrading quality, so we reject the new text if it loses in both rounds at each step.

We filtered out low-quality watermarked examples (e.g. those with a great number of repetitions (Zhao et al., 2023a)) since our perturbation oracle may persist in that repetition and the original text would not pass quality oracle in the first place. This can be a reasonable intervention as repetition violates our usable preassumption - we expect capable models like GPT-4 and future LMs not to produce repetitiveness after watermarking their outputs.

Stopping conditions. Users are allowed to design the stopping condition according to their needs and understanding of the detection: For the three watermarks we considered, we record and score each intermediate example and early stop on the one with z-score less than 1.645, which is prac-

tical whenever users roughly know the detection z-score threshold to make the stopping iteration a tunable hyper-parameter; For high-stake settings where users don't want to be detected as using AI-generated texts when users don't have any knowledge about the watermark scheme and detector or when the watermark is not robust enough (for long texts etc such as EXP (Kuditipudi et al., 2023), Unigram (Zhao et al., 2023a)), we stop when at least α of the words are replaced, where α is set to be 70% for texts of length 500 or 512, and proportionally for other lengths, e.g. 80% for length 600.

Table 3: The trade-offs of different quality oracle instantiation.

Model	Quality	(API) Costs	Efficiency
Reward Model	—	✓	✓
Reward Model + API Error Checking	✓	—	✓
GPT-3.5/4 API	✓	✗	✗

J.2 Watermark Details

Denote $|x|_G$ as the number of green list tokens for a generated text with length T. We experiment with three popular watermark schemes with their default hyper-parameters in general (Table 4).

- UMD (Kirchenbauer et al., 2023a) is about selecting a randomized set of "green" tokens before a word is generated, and then softly promoting the use of green tokens during sampling, which can be detected efficiently. We adopt a *one proportion z-test*, where $z = 2\left(|x|_G - T/2\right)/\sqrt{T}$. to evaluate the null hypothesis H_0:*The text sequence is generated with no knowledge of the red list rule* and choose to reject the null hypothesis if $z > 4$.
- EXP Kuditipudi et al. (2023) is a distortion-free watermark framework that preserves the original LM's text distribution, at least up to some maximum number of generated tokens. For detection, we compute a p-value with respect to a test statistic that measures the minimum cost alignment between length k subsequences of the text and key, via a permutation test with 5000 resamples. If ϕ returns a small p-value then the text is likely watermarked.
- Unigram (Zhao et al., 2023a) is proposed as a watermark robust to edit property. We calculate the number of green list tokens $|x|_G$ as well as the z-statistic $z = \left(|x|_G - \gamma T\right)/\sqrt{T\gamma(1-\gamma)}$ where $\gamma = 0.5$ means the fraction of the vocabulary included in the green list. The text is predicted as AI-generated if $z > 6$. We set the strength parameter $\delta = 2$, the larger δ is, the lower the quality of the watermarked LM, but the easier it is to detect.

Table 4: Default hyperparameters of our attack.

	UMD	EXP	Unigram
Attack steps	200	300	300
Secret key	15485863	42	0
z stopping threshold		1.645	
Max watermarked length		{200, 512}	
top-p of P		0.95	
Span length		{4, 6, 8}	
Num of spans l		1	
Min infill length		l	
Max infill length		$1.5 \times l$	

J.3 Prompt Examples

Recall that it is a design choice to implement the quality oracle as a combination of a reward model and GPT-3.5/GPT-4 to make sure the newly generated sample satisfies certain rubrics by rejecting samples with errors. In our case, we use the prompt in Figure 8.

Prompt for avoiding basic errors that can degrade text quality

Below are two candidate responses to the query {query Q}:
Original response: {response A}
New response: {response B}
Text quality is affected by factors such as unnecessary repetitions, grammar, coherence, relevance, and accuracy of the responses. Especially, having grammatical errors, repetitiveness, capitalization errors or punctuation mistakes would greatly degrade the quality of a response. Therefore, is the new modified response of equal or higher quality compared to the original response? If so, answer Yes, otherwise answer No.

Figure 8: GPT-3.5 prompt for avoiding basic errors that can degrade text quality.

We showcase our hand-crafted prompts for our GPT-3.5 or GPT-4 quality oracle for final evaluation (Figure 9). We also experiment with the system prompt in (Zheng et al., 2023), hoping to reduce biases regarding position, length etc but find that it has very minor effects on the results.

Prompt for evaluating and comparing text quality with five choices.

Below are two candidate responses to the query {query Q}:
Response A: {response A}
Response B: {response B}
Compare which of the two responses above is a better response to the given query. Explain your reasoning step by step.
(1) Response A is much better than response B
(2) Response A is slightly better than response B
(3) Responses A and B have similar quality
(4) Response B is slightly better than response A
(5) Response B is much better than response A

Figure 9: GPT-4 prompt for evaluating and comparing text quality with five choices

Figure 10: Intermediate text after attack (left, red) and its original watermarked text (right, green)

"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 stood out, even if he never fit in. He is the kid who never wanted 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, "even after coming up slowly all my life." He is the guy who's not afraid to speak his mind.

1 "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.

(a) The 5-th step, z-score = 5.35, p-value = 0.00, GPT-4 quality oracle score = 0.00.

"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 stood up when nobody else would. He is the kid who never wanted to sit on the bench, even if he didn't know if he'd ever get in a game. He is the kid who drove from Iowa City, Iowa, straight down the I-35 corridor to Indianapolis this week, so he could be part of the group. He is the kid, who when asked to sit down, he instead stood up. "I'm not a benchwarmer," Butler told The Athletic, "even after coming up slowly all my life." He is the guy who's not afraid to speak his mind.

1 "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.

(b) The 10-th step, z-score = 4.59, p-value = 0.00, GPT-4 quality oracle score = 0.00.

"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's Scouting Combine. They talk about technique, about routes, about film breakdowns. They show off their feet. They talk about their games. Butler is the outsider. He is the kid who always stood up when nobody else would. He is the kid who never wanted to sit on the bench, even if he didn't know if he'd ever get in a game. He is the kid who drove from Iowa City, Iowa, straight down the I-35 corridor to Indianapolis this week, so he could be part of the group. He is the kid, who when asked to sit down, he instead stood up. "I always stand out," Butler told The Athletic, "even after coming up slowly all my life." He is the guy who talks about his game and speaks his mind.

1 "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.

(c) The 15-th step, z-score = 2.78, p-value = 0.0027, GPT-4 quality oracle score = 0.00.

"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's Scouting Combine. They talk about technique, about routes, about film breakdowns. They talk about chemistry. They talk about their games. Butler is the outsider. He is the kid who always stood up when nobody else would. He is the kid who never, ever, wanted to sit on the bench, even if he didn't know if he'd ever get in a game. He is the kid who drove from Sioux City, Iowa, straight down the I-35 corridor, and to Indianapolis, so he could be part of the group. He is the kid, who when asked to step away, does so with conviction. "I always stand out," Butler told The Athletic, "even after coming up slowly all my life." He is the guy who talks about his game and speaks his mind.

1 "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.

(d) The 20-th step, z-score = 1.66, p-value = 0.048, GPT-4 quality oracle score = 0.00.

"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's Scouting Combine. They talk about technique, about routes, about film breakdowns. They talk about chemistry. They talk about their games. Butler is the outsider. He is the kid who always stood up when nobody else did. He is the guy who never, ever, wanted to sit on the bench, even if he didn't know if he'd ever get picked. He is the kid who drove from Sioux City, Iowa, straight down the I-35 corridor, just to be here, to join the rest of the group. He is the kid, who when asked to step away, does so with conviction. "I always stand out," Butler told The Athletic, "even after coming up slowly all my life." He is the kid, who talks about his game and speaks his mind.

1 "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.

(e) The 25-th step, z-score = 1.29, p-value = 0.0993, GPT-4 quality oracle score = 0.0

"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 Combine. They talk about their technique, about routes, about film breakdowns. They talk about chemistry. They talk about their games. Butler is the outsider. He is the kid who always stood up when nobody else did. He is the guy who never considered sitting on the end of the bench, even if he didn't know if he'd ever get picked. He is the kid who drove from Sioux City, Iowa, straight down the I-35 corridor, just to be here, to join the rest of the group. He is the kid, who when asked to step away, does so with conviction. "I always stand out," Butler said. "I worked my way up slowly all my life." He is the kid who talks about his game and speaks his mind.

1 "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.

(f) The 30-th step, z-score = 0.654, p-value = 0.257, GPT-4 quality oracle score = 0.00.

"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. They talk about their games. Butler is the outsider. He is the kid who always stood up when nobody else did. He is the kid who never considered sitting on the end of the bench, even if he didn't know if he'd ever make it off it. He is the kid who drove from Sioux City, Iowa, straight down the I-35 corridor, just to be here, to join the rest of the group. He is the kid who, when asked to step away, does so with conviction. "I always stand out," Butler said. "I worked my way up slowly all my life." He is the kid who talks about his game and speaks his mind.

1 "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.

(g) The 35-th step, z-score = 0.13, p-value = 0.448, GPT-4 quality oracle score = 0.00.

"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, stays out. He is the outsider. He is the kid who always stood up when he saw an opportunity. He is the kid who never considered sitting on the end of the bench, even if he didn't know 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 join the rest of the group. He is the kid who, when asked to step away, does so with conviction. "I always stand out," Butler said. "I worked my way up slowly all my life." He is the kid who talks about his game and speaks his mind.

1 "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.

(h) The 40-th step, z-score = 0.52, p-value = 0.303, GPT-4 quality oracle score = 0.00.

"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, stays out. He is the outsider. 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 happen. He is the kid who, when asked to step away, does so with conviction. "I always stand out," Butler said. "I worked my whole life." He is the kid who talks about his game and speaks his mind.

1 "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.

(i) The 45-th step, z-score = -0.26, p-value = 0.602, GPT-4 quality oracle score = 0.00.

"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.

1 "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.

(j) The 50-th step, z-score = 0.52, p-value = 0.30, GPT-4 quality oracle score = 0.00.