A Paraphrase Attack Safe and Distortion Free Watermarking Technique for AI-Generated Text

Anonymous ACL submission

Abstract

A report from the European Union Law Enforcement Agency forecasts that by 2026, up to 90% of online content may be synthetically generated (EUROPOL, 2022). This surge raises significant concerns among policymakers, who warn that "Generative AI could act as a force multiplier for political disinformation. The combined effect of generative text, images, videos, and audio may surpass the influence of any single modality" (Janjeva et al., 2023). In response, California's Bill AB 3211 mandates the watermarking of all AI-generated content (california legislature, 2023). However, existing watermarking techniques remain vulnerable to tampering and can potentially be circumvented by malicious actors. With the widespread adoption of Large Language Models (LLMs) across various applications, there is an urgent need for robust text watermarking solutions. Early watermarking models for LLMs proposed by Kirchenbauer et al. (2023a) faced criticism after studies by Sadasivan et al. (2024) and Chakraborty et al. (2023) demonstrated that paraphrasing could effectively remove these watermarks. In this paper, we introduce PECCAVI, the first text watermarking technique that is both resistant to paraphrase attacks and distortion-free, surpassing all existing methods in performance.

PECCAVI - at-a-glace

- Introducing a robust and flexible "*text watermarking framework*", designed for NLP applications, capable of marking and detecting model-generated outputs with high reliability. (cf. Sec. 3).
- We design an end-to-end *"sentence-level watermarking pipeline"*, integrating syntactic and semantic constraints to generate paraphrases with minimal distortion while preserving watermark detectability. (cf. Sec. 3).
- We implement a dynamic scoring system to "rank watermarked sentences "by balancing fluency, semantic similarity, and watermark strength, ensuring optimal sentence selection during inference. (cf. Sec. 4).
- We evaluate 150 watermarked sentences against 10 paraphrases, each to identify the most effective strategy, using a multi-metric ranking system and validating results across multiple paraphrase datasets. (cf. Sec. 4).

1 Introduction - the Necessity & Urgency

With the rise of prolific ChatGPT, the risk and consequences of AI-generated text has increased alarmingly. However, this rapid evolution and widespread accessibility presents significant challenges, particularly concerning the misuse of AI-generated images. In March 2023, an open letter (Marcus and of Life Institute, 2023) signed by numerous AI experts and industry leaders called for a six-month halt on the development of AI systems more advanced than GPT-4. The central concern noted in the letter (Marcus and of Life Institute, 2023) is "*Should we let machines flood our information channels with propaganda and untruth?*". While individual viewpoints on the notion of a moratorium may vary, the raised concern cannot be ignored. The findings of the latest (7^{th}) evaluation of the European Commission's Code of Conduct (European Commission, 2025) that seeks the eradication of mis/dis-information online reveals a decline in companies' responsiveness. The percent-

		z-score	p-value
Prompt	*Will the next great writer be a robot?*		
Watermarked text	I'm very skeptical that the next "great writer" is going to be a robot, or that they'll be much more effective at expressing the subtleties and depths of human thought than a human is. However, what is most interesting is the role that the Internet could play in bringing these "robot" writers into the public eye. If I could (and I'm very excited by this possibility), I would pay a small monthly fee to read well-written ...	4.24	1.1×10^{-5}
De-Watermarked text by replacing high-entropy words	I'm somewhat skeptical that the next "great writer" is going to be a robot, given that they'll be far more effective at grasping deeper subtleties and depths of philosophical thought than a robot is. However, what is particularly interesting is the role that the Internet may play in bringing new great writers into the public eye. If I did (and I'm extremely excited by this possibility), I would pay a hefty subscription fee to publish something ...	1.76 (58.5% ↓)	0.039 (3.5×10^{5}% ↑)
De-Watermarked text by paraphrasing	I have serious doubts about the possibility of a robot becoming the next exceptional writer and surpassing humans in expressing the nuances and profoundness of human thoughts. Nevertheless, what fascinates me the most is the potential impact of the Internet in showcasing these "robot" writers to the general public. The idea of being able to pay a nominal monthly subscription fee to access impeccably written and carefully refined works truly thrills me...	-0.542 (112.8% ↓)	0.706 (6.4×10^{6}% ↑)

Table 1: An illustration of de-watermarking by replacing high-entropy words and paraphrasing. p-value is the probability under the assumption of null hypothesis. The z-score indicates the normalized log probability of the original text obtained by subtracting the mean log probability of perturbed texts and dividing by the standard deviation of log probabilities of perturbed texts. DetectGPT (Mitchell et al., 2023) classifies text to be generated by GPT-2 if the z-score is greater than 4.

age of notifications reviewed by companies within 24 hours decreased, falling from 90.4% in 2020 to 64.4% in 2022. This decline likely reflects the increased accessibility of Gen AI models, leading to a notable influx of AI-generated content on the web.

To tackle the misuse of AI-generated content, two primary approaches have emerged: *pre-hoc* and *post-hoc* techniques. Watermarking is a pre-hoc method that involves embedding detectable markers into AI-generated content. For watermarking to be effective, it must be adopted by LLM and Gen AI model providers and could become widespread only if mandated by government regulations. In the absence of such regulations, open-source LLMs and Gen AI models without watermarking continue to proliferate, resulting in a surge of AI-generated content online. Therefore, there is a pressing need for automated post-hoc techniques to identify whether text found online was generated by AI.

Governments worldwide have begun discussions and have implemented measures to develop policies concerning AI systems. The European Union (European-Parliament, 2023) has taken a definitive stance by enacting legislation, while the United States (White-House, 2023) and others have introduced preliminary proposals regarding the regulatory framework for AI. One of the primary concerns among policymakers is that "*Generative AI could act as a force multiplier for political disinformation. The combined effect of the generative text, images, videos, and audio may surpass the influence of any single modality*" (Janjeva et al., 2023). Additionally, AI policymakers[1] have raised significant concerns about the use of automatic labeling or invisible watermarks as a technical solution to the challenges posed by generative AI-enabled disinformation. However, there are ongoing apprehensions about the susceptibility of these measures to deliberate tampering and the potential for malicious actors to bypass them entirely.

2 Paraphrase Attack Safety

A core idea behind proposing the text watermarking technique, is the manipulation of high-entropy words, content-rich terms that contribute significantly to the semantics of a sentence. These words are selectively replaced with contextually plausible alternatives, chosen by an algorithm that functions

[1] https://cetas.turing.ac.uk/publications/rapid-rise-generative-ai

similarly to an *encryption key*, known only to the LLM creator. This design choice stems from the distinction between high- and low-entropy words in natural language. High-entropy words carry the primary informational content, whereas low-entropy words (e.g., prepositions, conjunctions) mainly serve syntactic and structural roles. Replacing low entropy words can disrupt the quality of text generation. (Bentz et al., 2017) provides more details on high-entropy vs. low-entropy words.

2.1 Related Works - Absence of Fully Secure Paraphrase Attack Safety

The detection of AI-generated text has been a burning topic of research for a long time now, especially after the exponential growth of Large Language Models in daily users' lives, leading to the development of various watermarking techniques designed to verify text authenticity. A few notable approaches were proposed earlier by Hou, Abe B. et al. (2023) and Dathathri et al. (2024).

SemStamp - A Semantic Watermark with Paraphrastic Robustness for Text Generation: Traditional token-level watermarking techniques (Kirchenbauer et al., 2023b) rely on token distribution statistics, making them vulnerable to paraphrasing (Chakraborty et al., 2023). To address this, (Hou et al., 2024a) proposed *SemStamp*, a sentence-level watermarking method that embeds sentences into a semantic space and partitions it using locality-sensitive hashing (LSH) (Indyk and Motwani, 1998) or, in a later refinement, k-means clustering (Hou et al., 2024b).

Watermarked regions are defined within this space, and rejection sampling is used during generation to ensure outputs fall within those regions. Detection involves embedding the sentence and checking if it lies in a watermarked partition. While more robust than token-level methods, *SemStamp* remains susceptible to inter-sentence paraphrasing and loses reliability when watermarked content is mixed with human-written text.

SynthID-Text - Scalable Watermarking for Identifying Large Language Model Outputs: *SynthID-Text*, introduced by (Dathathri et al., 2024), propose a generative watermarking technique that operates in a pre-hoc manner. This technique integrates a random seed generator, a sampling algorithm, and a scoring function. The language model generates candidate tokens, each token is then assigned scores using multiple random functions, and the candidate with the highest score is sampled by the tournament sampling approach.

3 $\mathcal{PECCAVI}$: A Paraphrase Attack Safe and Distortion Free Watermarking Technique for AI-Generated Texts

PECCAVI design addresses the vulnerability of watermarks to paraphrase attacks by systematically simulating these transformations to identify optimal embedding locations. The study follows a structured approach, beginning with the identification of Non-Melting Points (NMPs), which remain stable across paraphrases. Once determined, n masking methods are applied to mask specific words or word groups, and the outputs are processed using m sampling methods, yielding $n \times m$ unique combinations. Each variation is then paraphrased 10 times to evaluate watermark distortion and detectability based on the average changes observed. This process is repeated for every sentence in the dataset, ultimately identifying the most effective balance between high detectability and low distortion.

3.1 Dataset

The dataset used in this study is the NY Times dataset, which consists of a diverse collection of news articles and tweets. This dataset provides a rich and varied textual corpus, ensuring a robust evaluation of watermarking techniques across different writing styles and topics.

3.2 Where to Add the Watermark?

Determining **where** to insert a watermark is crucial in text watermarking. Traditional approaches,

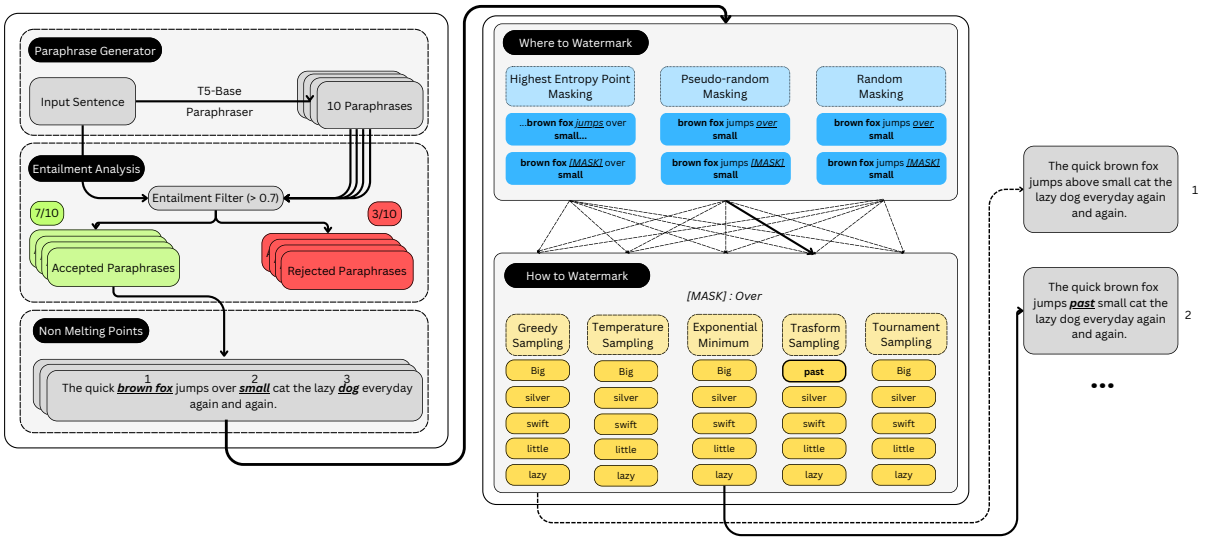

Figure 1: Illustration of PECCVAI Text section 1. The Paraphrase Generator (left) produces paraphrases, which undergo Entailment Analysis, filtering with some threshold. The Watermarking Process (right) consists of "Where to Watermark"—using Highest Entropy Point Masking, Pseudo-random Masking, or Random Masking—and "How to Watermark", which applies different sampling methods (Greedy, Temperature, Exponential Minimum, Transform, and Tournament Sampling) to replace masked words. This method embeds detectable watermarks while preserving text coherence.

such as embedding in high-entropy words, are fragile against simple attacks like word replacement and ineffective against sentence-level paraphrasing as illustrated in Figure 1. To address these limitations, the proposed method identifies **Non-Melting Points (NMPs)** words or N-grams that remain consistent across paraphrased versions of a sentence. At least 10% of the words in a sentence should qualify as NMPs, serving as reference points for selecting watermark locations. Words between consecutive NMPs are identified as potential candidates, and three different masking techniques are applied to determine the optimal embedding strategy while preserving robustness and readability.

In the example sentence: *The quick brown fox leaps over a small cat and a lazy dog every day again and again*, the identified NMPs are "brown fox," "small cat," and "lazy dog." The words between these NMPs are potential watermarking candidates.

In **Highest Entropy Masking**, the word with the highest entropy between consecutive NMPs is masked, ensuring that the most unpredictable and meaningful words are altered. This enhances watermark detectability and robustness against substitutions. Applied to the example, the sentence transforms into: *The quick brown fox [MASK] over a small cat and a lazy dog [MASK] day again and again*, where "leaps" and "every" are masked due to their high entropy.

In **Random Masking**, a word between two NMPs is selected at random, introducing variation and making the watermarking pattern less predictable while preserving readability. Applied to the example, the sentence transforms into: *The quick brown fox leaps over a small [MASK] and a lazy dog every day again and again*, where "cat" was randomly chosen for masking.

In **Pseudo-Random Masking**, a fixed seed ensures consistent word selection across runs while remaining unpredictable to attackers. Applied to

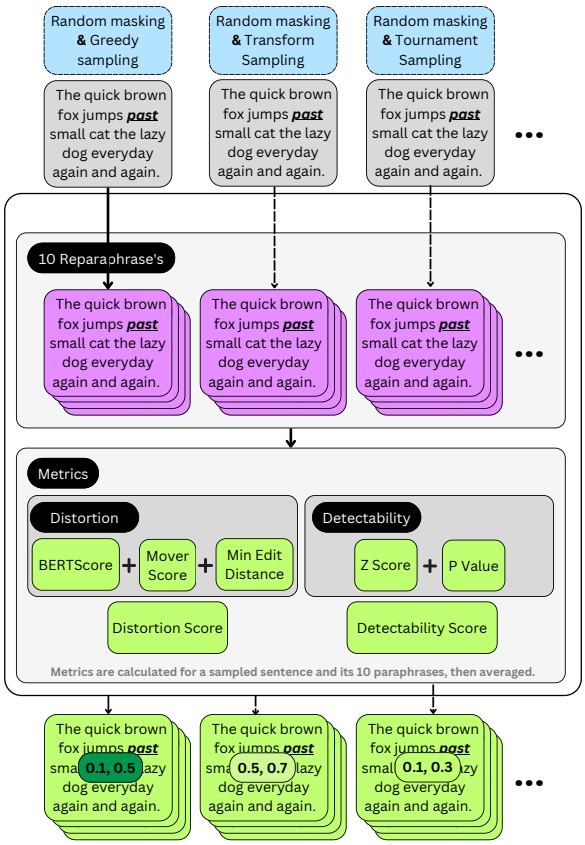

Figure 2: Illustration of PECCVAI Text Section 2. The re-paraphraser generates 10 paraphrases per sentence, evaluated using distortion and detectability metrics. Distortion includes BERTScore, Mover Score, and Edit Distance, forming the Distortion Score. Detectability includes Z-score and P-value, forming the Detectability Score. The metrics section shows example paraphrases with minor lexical and structural variations, highlighting how watermarked text remains detectable while preserving meaning. Metrics are averaged across paraphrases.

the example, the sentence transforms into: *The quick brown fox leaps over a small cat and a lazy [MASK] every day again and again*, where "dog" was masked based on the pseudo-random selection process.

Selecting multiple words for watermarking ensures that multiple watermarks are embedded within the text, increasing robustness. These masking techniques enhance resistance to common attacks like word replacement and sentence paraphrasing while preserving readability and coherence.

3.3 Watermarking Strategies

Once the target word or word group is selected for watermarking, the choice of **how** to embed it is equally crucial. The effectiveness of a watermarking technique depends not only on *where* it is inserted but also on the *how*, making sampling methods essential. Sampling in text watermarking involves selecting the optimal replacement or modification while preserving readability and meaning. A systematic approach ensures detectability while minimizing distortion and enhancing resistance to attacks like paraphrasing, synonym replacement, or adversarial transformations.

Greedy sampling selects the candidate with the highest probability at each step, ensuring consistent and predictable watermark placement. Example output: *The quick brown fox jumps over the lazy dog*

Temperature sampling adjusts randomness using a temperature parameter. Higher temperatures increase variation, while lower ones make selection more deterministic, balancing predictability and diversity. Example output: *The fast brown fox leaps over the lazy dog*

Exponential minimum sampling favors low-probability candidates by applying an exponential transformation, often selecting rare but plausible substitutions. Example output: *The agile brown fox soars over the lazy dog*

Transform sampling applies a nonlinear transformation to probability scores, adapting to sen-

tence structures and enhancing contextual integration. Example output: *The speedy brown fox vaults over the lazy dog*

Tournament sampling selects a random subset of candidates before choosing the best option, combining randomness with robustness. Example output: *The swift brown fox hurdles over the lazy dog*

By incorporating these sampling methods, our watermarking framework ensures a balance between detectability, robustness, and linguistic coherence. The combination of different strategies allows for flexibility in watermark placement, making it adaptable to various types of textual data and resistant to common attacks.

4 Performance of Watermarking

4.1 Detection and Distortion Metrics

The effectiveness of a text watermark is determined by two key factors: distortion and detectability. Distortion is the degree to which a paraphrased sentence deviates from the original sentence. Here the deviation can be in terms of lexical, syntactic, and semantic properties. A higher level of distortion can lead to loss of factual accuracy and alteration of the original idea or meaning.

Detectability assesses how easily the watermark can be identified by a detector.

Distortion is evaluated using a combined score based on three metrics:

Minimum Edit Distance: Levenshtein distance or minimum edit distance is a metric that quantifies the effort required to paraphrase and it does that by keeping track of elementary operations like insertions, deletions, and substitutions required to convert *sentence 1* to its paraphrase *sentence 2*. Levenshtein distance is a good metric to calculate the surface-level distortion but it does not take semantic similarity into consideration - a paraphrase with different words but the same meaning can have a high Levenshtein distance which can signal a higher distortion, however, the meaning may still be retained along with the factual accuracy.

BERTScore: To capture the semantic similarity between the original sentence and the paraphrased versions, the study used BERTScore (Zhang et al). BERTScore utilises the contextual embeddings from transformer-based language models like BERT to calculate the similarity between any two sentences. Unlike Levenshtein distance which relied on calculating the exact difference between the string pairs, BERTScore calculates token-wise cosine similarity between the embeddings of tokens in the *sentence 1* and *sentence 2*.

Mover Score: Captures semantic differences by measuring similarity between the original and watermarked text. MoverScore builds on Word Mover's Distance by measuring the semantic distance between two sentences using word embeddings. It uses cosine similarity to assess closeness between word pairs and applies the Hungarian algorithm for optimal word matching. Additionally, TF-IDF weights are used to prioritize important words. The result is a TF-IDF-weighted similarity score, higher MoverScore means higher similarity. In our case, since we focus on dissimilarity, we compute it as: Dissimilarity = 1 – MoverScore

The final distortion score is obtained by averaging these metrics over 10 paraphrased versions of the original sentence and the sampled watermarked sentence.

Detectability is evaluated using a combined score based on two statistical methods, specifically the **Z-score** and **P-value**: The Z-score indicates how significantly the watermark deviates from the expected distribution, while the P-value determines whether the observed difference is statistically significant.

4.2 Results

4.3 Which watermarking technique works best?

The watermarking strategy which offers lower distortion and at the same time higher detectability is the optimal watermarking strategy. Lower distortion means that the watermarked sentence does not

	Random Masking		Pseudo Random Masking		Highest Entropy Masking	
	Distortion	Detection	Distortion	Detection	Distortion	Detection
Greedy Sampling	0.53	0.81	0.54	0.85	0.46	0.88
Temperature Sampling	0.5	0.82	0.63	0.91	0.56	0.83
Tournament Sampling	0.49	0.84	0.49	0.84	0.46	0.89
Transform Sampling	0.57	0.90	0.51	0.84	0.48	0.77
Exponential Minimum Sampling	0.52	0.82	0.53	0.84	0.47	0.79

Table 2: Comparison of Masking Methods with Sampling Techniques, including Distortion and Detectability

change much in meaning compared to the original sentence and high detectability means that the sentence still shows signs of watermarking after it is paraphrased. The distortion vs detectability results are shown in 2

5 Conclusion

With the rapid proliferation of AI-generated text, ensuring content authenticity has become a critical challenge. Existing watermarking methods remain vulnerable to paraphrase attacks, where minor rewording can effectively erase embedded markers. In response, we introduced PECCAVI, a novel watermarking technique designed to be both paraphrase attack-resistant and distortion-free, ensuring watermark retention while maintaining linguistic coherence. Our approach leverages Non-Melting Points (NMPs) to identify stable embedding locations and applies a combination of entropy-based masking and adaptive sampling to generate robust watermarks. Through extensive experimentation, we demonstrated that PECCAVI outperforms existing watermarking techniques, offering a higher degree of resilience against paraphrasing while preserving the readability of AI-generated content.

6 Discussion and Limitations

Overlapping boxes, distorts the wm pattern, computation time **Discussion:** On June 14th, 2023, the European Parliament successfully passed its version of the EU AI Act (European-Parliament, 2023). Following this, many other countries began discussing their stance on the evolving realm of Generative AI. A primary agenda of policymaking is to protect citizens from political, digital, and physical security risks posed by Generative AI. While safeguarding against misuse is crucial, one of the biggest concerns among policymakers is the occurrence of unwanted errors by systems, such as hallucination (source: https://cetas.turing.ac.uk/publications/rapid-rise-generative-ai).

Limitations: The empirical findings indicate that classifying temporal issues poses the greatest challenge, as shown in Figure **??**. (Gurnee and Tegmark, 2023) claimed that LLMs acquire linear representations of space and time across various scales, it is expected that LLMs hold such information internally and can classify accordingly. Performance on temporal issue 66% is not bad but could be seen as a future direction to improve.

Despite its strengths, PECCAVI is not without limitations. Its computational overhead remains a concern, particularly for real-time applications, and its effectiveness across diverse LLM architectures warrants further evaluation. Additionally, evolving adversarial techniques may necessitate continuous refinements to improve robustness against more sophisticated attacks.

Looking ahead, several future directions can further enhance the applicability and effectiveness of PECCAVI. Key areas for exploration include optimizing computational efficiency, improving cross-model generalization, and integrating hybrid watermarking approaches that combine statistical and semantic techniques for greater robustness. Moreover, policy-level considerations will play a crucial role in encouraging the adoption of watermarking solutions within AI governance frameworks.

By addressing these challenges, PECCAVI has the potential to become a standard for AI text authentication, ensuring transparency and security in an era where generative AI continues to reshape digital content creation.

7 Ethical Considerations

Through our experiments, we have uncovered the susceptibility of LLMs to hallucination. While emphasizing the vulnerabilities of LLMs, our goal is to underscore their current limitations. However, it's crucial to address the potential misuse of our findings by malicious entities who might exploit AI-generated text for nefarious purposes, such as designing new adversarial attacks or creating fake news that is indistinguishable from human-written content. We strongly discourage such misuse and strongly advise against it.

References

Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i Cancho. 2017. The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy*, 19(6):275.

california legislature. 2023. Ab-3211 california digital content provenance standards.

Megha Chakraborty, S.M Towhidul Islam Tonmoy, S M Mehedi Zaman, Shreya Gautam, Tanay Kumar, Krish Sharma, Niyar Barman, Chandan Gupta, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. 2023. Counter Turing test (CT2): AI-generated text detection is not as easy as you may think - introducing AI detectability index (ADI). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2206–2239, Singapore. Association for Computational Linguistics.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Merey, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, Ilia Shumailov, Ciprian Baetu, Sven Gowal, Demis Hassabis, and Pushmeet Kohli. 2024. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823.

European Commission. 2025. 7th evaluation of the code of conduct on countering illegal hate speech online. Accessed: 2025-02-11.

European-Parliament. 2023. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.

EUROPOL. 2022. Facing reality?: Law enforcement and the challenge of deepfakes.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *Preprint*, arXiv:2310.02207.

Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. 2024a. Semstamp: A semantic watermark with paraphrastic robustness for text generation. *Preprint*, arXiv:2310.03991.

Abe Bohan Hou, Jingyu Zhang, Yichen Wang, Daniel Khashabi, and Tianxing He. 2024b. k-

semstamp: A clustering-based semantic watermark for detection of machine-generated text. *Preprint*, arXiv:2402.11399.

Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, page 604–613, New York, NY, USA. Association for Computing Machinery.

Ardi Janjeva, Alexander Harris, Sarah Mercer, Alexander Kasprzyk, and Anna Gausen. 2023. The rapid rise of generative ai: Assessing risks to safety and security.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023a. A watermark for large language models. In *International Conference on Machine Learning*.

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

Gary Marcus and Future of Life Institute. 2023. Pause giant ai experiments: An open letter.

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

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2024. Can ai-generated text be reliably detected? *Preprint*, arXiv:2303.11156.

White-House. 2023. Blueprint for an ai bill of rights: Making automated systems work for the american people.

[