Robustness of LLM Text Watermarking: A Critical Review of Evasion Vectors and Defense Mechanisms with a Hybrid Semantic–AMR Case Study

M. Nanda Krishna[1],Dr V. Sujatha Lakshmi[2]

[1]Independent Researcher, Bhashyam Junior College, Guntur, India.

[2]Associate Professor,Department of CSE, RVR&JC College of Engineering, Chowdavaram,Guntur.

[*]Corresponding Author

DOI: https://doi.org/10.51244/IJRSI.2026.1311000060

Received: 05 November 2026; Accepted: 10 November 2026; Published: 27 November 2026

ABSTRACT

The proliferation of LLM use in society makes the problem of detecting machine-written text from human text a matter of academia and an important issue to reduce automated mis-information. One method of watermarking generated text is statistical or cryptographic text watermarking; however studies recently have been published that suggest watermarking be-comes ineffective against a malicious transformation. In this paper, this author compiles the bibliography summarizing the performance of the watermarking methods for LLM (literature) and classifies the threats discussed in the literature into two main categories: evasion attacks (paraphrasing, translating, rewriting) and spoofing attacks (token level modification, false at-tribution). It also contrasts the types of detection and evaluation which these studies employ, highlighting the use of z-scores for detection, true positive rate (TPR), false positive rate (FPR), an area under the ROC curve (AUC), F1 score, token requirement, and attack success. The experimental results of a supplementary experiment are finally compared using a token-level (KGW) baseline and semantic and Abstract Meaning Representation (AMR)-based and hybrid watermark architectures in four attack scenarios, with the hybrid architecture exhibiting the highest performance in both AUC and TPR. The watermark strength versus the reliability of its detection, the quality or robustness of the text, and the ability to withstand semantic-preserving transformations continue to tug at literary writing.

Keywords: Large Language Models, AI Watermarking, Adversarial Machine Learning, Evasion Attacks, Cyber security, Robustness, Detection Metrics

INTRODUCTION

The performance of Large Language Models (LLMs) has dramatically changed the way textual information is created, shared, and interpreted in many fields from education to scientific research, digital communication to journalism, health to automated content generation. Modern genera-tive language models can be used to generate contextually relevant, grammatically correct and semantically meaningful language with little to no human effort. While these features enhance pro-ductivity and access, they also bring significant challenges in terms of content verification, spreading misinformation, protecting intellectual property, academic integrity, and copyright infringement of machine-generated content. The problem of distinguishing between human-made and AI-generated text has therefore become a significant area of research in the field of trustworthy AI. This research has recently highlighted the need for trustworthy mechanisms for content-provenance that must correctly determine if content has been generated by a machine, while preserving its linguistic quality and also limiting the use of generative models for legitimate purposes [1]–[3].

A novel method for establishing a digital provenance of LLM output text has been proposed, whereby a statistical, cryptographic or semantically identifiable signature is embedded into the text generated by the LLM. While the traditional post-hoc AI-text classification methods include adding a detectable signal to the generated content, watermarking is done intentionally by embedding a signal within the process of creation or a carefully manipulated transformation of the final output. The resulting text can then be analyzed by an authorized detector to check for the presence of the desired watermark signal. In an influential token-level watermarking framework, designed by Kirchenbauer et al. [1], the vocabulary is split into two pseudorandom subsets of green and red tokens, and the probabilities of selecting tokens are changed to form statistically detectable patterns. They have developed a feasible basis for generation-time watermarking by hypothesis testing and interpretable detection statistics. Later research concentrated on furthering the watermark design in the direction of reliability, robustness, and resistance to typical watermark modifications found in real-world text processing [2]–[4].

However, although these advances have been made, the strength of text watermarking continues to be a challenge due to the fact that many different natural expressions can convey the same message. Paraphrasing, translating, substituting synonyms, restructuring sentences, summarizing, deleting sentences, and recursive rewriting watermarked content may change the meaning of what was originally intended, as long as no significant changes are made. These changes may affect the token frequency, sequences of tokens, and the context of tokens within the text such that the token-level detectors are unable to recognize watermark signatures. Thus, a text sample successfully classified as watermarked soon after it's produced can be harder to detect after it's been subject to an adversarial modification. Research into semantic watermarking has shown that watermarks in higher-level linguistic properties can minimize the reliance on exact token identity, and thus can be resistant to certain meaning-preserving transformations [5]–[8]. However, the robustness still relies on the embedding mechanism, the semantic representation, the transformation intensity and the detection threshold.

Apart from watermark removal, the adversarial manipulation can also include watermark spoof-ing, false attribution and unauthorized signature reconstruction. The security threat of these at-tacks is different since an attacker could try to destroy evidence that the machine is generating, or the attacker could try to use unrelated text to make it seem like it has a valid watermark. Research on watermarks has confirmed that security levels can only be determined by the ability to successfully recognize the watermark and thus cannot be done exclusively by the recognition of unmodified watermark text. Rather, adversarial performances, watermark properties, text quality, detector calibration and potential wrong attribution should be taken into account [9]–[11]. The high detection performance in controlled settings does not necessarily guarantee security against adaptive attackers, and, for practical purposes, theoretical guarantees of watermark removal can rely on conditions that are hard to meet in real applications.

These concerns have been tackled recently by the use of a semantic-aware watermarking, op-timized token selection, distribution-preserving techniques and unified benchmarking protocols. Semantic watermarking methods like SemStamp and its clustering-based extension cluster the sen-tences generated based on semantic embeddings, rather than only on lexical token patterns [5], [6]. Some other methods adopt a contextual semantic representation, invariant linguistic features or post generation modification to enhance the persistence of the watermark against rewriting [3, 7, 8]. Meanwhile, a number of systematic optimization and benchmarking studies have highlighted the inherent trade-off between detectability, compute efficiency, generated-text quality, and adversarial robustness [9, 12–17].

One development of particular interest is the use of Abstract Meaning Representation (AMR) that encodes sentence meaning as a graph structure with concepts and their semantic relationships. AMR is a structured description of semantic roles and relationships, and it is less lexical than purely lexical representations, in that it may be somewhat consistent across sentences that have been rewritten with different words and phrases but have the same meaning. The SWAN framework shows the feasibility of embedding watermark signatures into AMR-based semantic structures and shows a crucial basis for

studying structurally informed watermark detection [18]. Other recent embedding space techniques also stress the overall significance of semantic invariance in watermarks [19]. These results call for research into complementary semantic and structural evidence sources, not on a single source.

However, a successful Hybrid watermarking system does not simply consist of a combination of multiple detector scores. There are various aspects of language that can be represented with different types of semantic embedding and AMR representations and adding them together results in further challenges related to normalization, evidence fusion, thresholding, computational load, and false-positive control. An embedding-based detector may have difficulty detecting the differences in AMR between paraphrases, while a structural detector can be sensitive to the difference in how the graph is constructed or AMR parsing errors. Some of these individual limitations could be overcome with a complementary detection strategy, which needs to be proven in controlled and reproducible experiments. These investigations could be supported by some helpful bases provided by recent studies on robust watermark generation [19]–[20] and semantic preservation [21]–[22] and systematic attack evaluation.

The other problem is that in the current watermarking studies, there is no fully standard exper-imental setup for the experiment. Existing methods often vary with respect to language models, prompt datasets, watermark embedding strength, prompt text length, attack implementation, de-tection threshold. Therefore, direct comparisons between results reported by different studies using the true positive rate (TPR), false positive rate (FPR), area under the receiver operating charac-teristic curve (AUC), F1-score or statistical significance tests may not always be possible. Bench-marking techniques include WaterBench, MarkLLM and WaterPark, each offering a way to create a more uniform evaluation process, comparable conditions and structured attack process [12]–[14]. Recent surveys also emphasize the importance of transparency in reporting, comprehensive threat modelling, and reproducible validation in assessing the effectiveness of LLM watermarking systems [16, 17, 21, 22]. Beyond watermark-specific benchmarks, related infrastructure work addresses se-cure middleware deployment and cloud workflow orchestration, both relevant operational concerns for reproducible evaluation pipelines [26], [27].

In response, the current research examines the resilience of LLM text watermarking by first reviewing adversarial evasion techniques and defenses and further analyzing an experimental setup leveraging a combination of Semantic and AMR watermarking techniques. Literature review clas-sifies watermarking methods based on their representation, and analyze the vulnerability of the methods to paraphrasing, translating, manipulating tokens, recursive rewriting, random walking editing, and false-attribution attacks. Four watermarking configurations are evaluated: token-level KGW, structural watermarking using semantic embedding, structural watermarking based on ad-vanced machine recognition (AMR) and a final watermarking configuration combining semantic and AMR evidence. Detection process is based on the watermark persistence after controlled text transformations and the semantic similarity between detected and reference images, AMR simi-larity, combined evidences from image detection and image classification performance. The major contribution of this work is a combined analysis of the performance of various watermark represen-

tations under different kinds of adversarial transformations, and of the complementary information that can be obtained from both the watermark semantic and structural content to aid watermark detection. Furthermore, the study highlights the text preservation-watermark preservation trade-off, examines how detector performance varies as a function of adversarial modification and helps to identify the methodologically needed properties for a statistically credible evaluation of watermark. The results of the experiments are not viewed as valid for universal superiority but instead as only a small sample study. By synergizing both the literature-based and experimental viewpoint, the researchers hope to help develop clear, reliable, and standardized assessment tools for trustworthy LLM-generated text identification.

LITERATURE SURVEY

Token-Level and Statistical Watermarking

In recent years, a watermarking technique at the token level has become an effective method for identifying the author, context, and/or origin of LLM-generated text data using statistically detectable token-selection patterns. Kirchenbauer et al. [1] proposed a green-list/red-list water-marking system, which alters the token probabilities, and uses z-scores for detection. In their next work [2] they investigated the reliability of watermarks when they are paraphrased or rewritten by human. They showed that the length of the text and the intensity of perturbation have an influ-ence. Yoo et al. [3] proposed watermarking based on two levels of invariants features to enhance the robustness against the textual manipulation while Christ et al. [4] studied on the cryptographically undetectable watermarking. A balance between watermark detectability and watermark generation quality was analysed by Wouters [9] and a robust watermark approach was proposed by Ren et al. [20] based on the concept of token prior probabilities. All these developments notwithstanding, much work on token-level techniques is still stymied by large lexical and contextual changes.

Semantic-Based Watermarking

The objective of semantic watermarking techniques is to enable removal of some text changes without destroying the watermark while still leaving the signature visible. Hu et al. [5] proposed SemStamp, a locality-sensitive hashing over sentence embeddings based method to enhance the robustness of paraphrasing. It is the extension of k-SemStamp [6] which uses clustering-based semantic partitioning to improve the detection reliability. SemaMark [7] is suggested by Ren et al., which is based on the semantic information to improve watermark robustness against para-phrasing attack. An embedding-space watermarking approach that can resist semantic-invariant attacks is presented by Ai and He [19] and a black-box watermarking method, named PostMark, is proposed by Chang et al. [8] that embeds signatures after the text generation. These methods leads to increased resistance to selected meaning preserving transformations; but their performance is sensitive to the quality of semantic representations, embedding stability and to parameters of the detection.

Color Correction; Darkening or Brightening.

Abstract Meaning Representation (AMR) is a graph-based semantic representation of sentence meaning, based on concepts and their relations. Ye et al. [18] proposed a new AMR-based water-marking technique named SWAN, which embeds structural signatures to enhance the robustness to paraphrasing attacks. The AMR-based methods take advantage of semantic relationships, which

can stay the same even if lexical changes occur. The reliability however requires semantic parsing accuracy and consistency of the structure. Using AMR-based structural evidence with semantic embeddings is a potentially useful approach for hybrid watermark detection. As a consequence, the proposed framework uses semantic similarity, AMR structural matching and sentence-level consistency to study the persistence of the watermark in adversarial transformations.

Adversarial Attacks and Watermark Security

Adversarial Transformations have a significant impact on the reliability of LLM watermarking systems. In this paper, Sadasivan et al. [23] studied the basic barriers surrounding adversarial rewriting detection within AI-generated text. An in-depth theoretical analysis of strong water-marking was proposed in Zhang et al. [10] and watermark stealing via scrubbing and spoofing attacks were shown in Jovanović et al. [11]. Liang et al. [14] proposed a comprehensive evalua-tion framework, called WaterPark, which evaluated several watermarking techniques and evaluated them in the face of representative attacks from watermarkers. Harel-Canada et al [15] went on further to evaluate the practical text quality constraints to random-walk watermark removal. The studies show the implanted watermarks have various degrees of vulnerability, like paraphrasing, translation, token deletion,

recursive rewriting, and false attribution, and clearly indicate that the robustness of the watermarks must be measured in addition to preserving the semantics and providing good generation quality.

Evaluate performance against expectations based on the results of bench-marking.

Standardized experimental procedures and consistent measures are needed to have reliable eval-uation. Tu et al [12] built WaterBench to evaluate the detectability of the watermark and text quality for uniform watermark strength. Hahn et al. [10] proposed a structure of an open-source watermarking toolkit suitable for systematic assessment and evaluation of watermarks, and Piet et al. [24] discussed the tamper resistance and detection capabilities. Liu et al. [16] and Lalai et al. [17] gave a summary of existing watermarking techniques, threats to security, and evaluation methods followed by Yang et al. [21] and Liang et al. [22] and finally Wang et al. [25]. These investigations highlight the need to report true positive rate (TPR), false positive rate (FPR), area under curve (AUC), statistical detection evidence and experimental repeatability. Yet, direct com-parisons remain problematic because of differences in datasets, watermark setup, choice of attack type and attack parameters, and choice of detection threshold.

Research Gap

The current LLM watermarking schemes show significant performance enhancements on detectabil-ity and adversarial transformations. Nevertheless, token-based methods suffer from lexical changes, semantic watermarking requires participation stability and structure-based methods require accu-rate structural representations. In addition, there is still limited literature about how well the two types of watermark signals can be combined consistently in an adversarial setting. Other assess-ments also adopt different experimental setups, statistics reporting and/or measures of robustness. Hence, the present research focuses on a hybrid Semantic–AMR watermarking framework combining semantic evidence, structural matching and sentence-level consistency that explores the robustness of the watermark against paraphrasing, translation, deletion (or ab trim) and random-walk at-tacks. Comparative experiments are provided with detection and robustness measures for the 3 configurations (KGW, semantic, AMR and hybrid) and a larger number of repeated experiments is necessary to prove statistical reliability.

Table 1 summarizes the principal methods, evaluation measures, and methodological limitations considered in this review. The reported metrics are drawn from the cited studies and should be interpreted in the context of each study's experimental setup.

Table 1: Comparison of LLM text watermarking methodologies, evaluation measures, and limitations.

Existing Methodology	References	Statistical Evaluation Measures	Methodological Limitations
KGW Token-Level Watermarking	[1, 2]	Z-score, p-value, TPR, FPR, detection threshold	Vulnerable to paraphrasing, translation, and token substitutions; detection depends on text length and token distribution.
Invariant-Feature Watermarking	[3]	Bit recovery accuracy, detection accuracy, robustness under corruption	Invariant features may change under extensive rewriting; multi-bit recovery can become unreliable.
Cryptographic Watermarking	[4]	Detection advantage, computational indistinguishability, security guarantees	Cryptographic security does not guarantee robustness against meaning-preserving rewriting.
SemStamp	[5]	Detection rate, FPR, semantic similarity, paraphrasing robustness	Sensitive to semantic partition boundaries and embedding variations; computationally expensive candidate

			filtering.
k-SemStamp	[6]	Detection performance, sampling efficiency, paraphrase robustness	Depends on clustering quality, embedding stability, and cluster-boundary selection.
SemaMark	[7]	Detection performance, paraphrasing robustness, text quality	Semantic representations may be affected by extensive structural rewriting or semantic drift.
PostMark	[8]	Detection rate, watermark robustness, semantic preservation	Post-generation modifications may affect textual naturalness and generation quality.
Optimized Watermarking	[9]	Detection–quality Pareto trade-off, detection effectiveness	Optimization outcomes depend on selected objectives, constraints, and model configurations.
Watermark Removal Analysis	[10]	Theoretical robustness and adversarial success conditions	Theoretical attack conclusions depend on assumptions about allowable modifications and quality preservation.
Watermark Stealing	[11]	Spoofing success, watermark scrubbing, adversarial effectiveness	Vulnerable to adaptive attackers and information leakage through repeated model queries.
WaterBench	[12]	TPR, FPR, detection performance, generation quality	Benchmark conclusions depend on selected language models, tasks, and matched watermark strength.
MarkLLM	[13]	Detectability, robustness, text quality, evaluation efficiency	Provides standardized evaluation tools but does not independently guarantee statistical reproducibility.
WaterPark	[14]	TPR, FPR, AUC, attack robustness	Performance varies across watermark algorithms, attack conditions, and operating thresholds.
Random-Walk Removal Analysis	[15]	Attack success rate, quality preservation, detector performance	Attack outcomes depend on edit budgets and the reliability of text-quality assessment.
AMR-Based SWAN	[18]	Watermark detection, structural robustness, paraphrase resistance	Performance depends on AMR parsing accuracy, structural consistency, and graph representation.
Embedding-Space PASA	[19]	Detection reliability, robustness, semantic preservation	Requires further validation across diverse models, languages, and adversarial transformations.
Proposed Hybrid Semantic–AMR	**Present study**	**AUC = 0.988; TPR = 94.2%; FPR = 0.9%**	Preliminary single-case comparison; repeated trials, confidence intervals, statistical significance tests, and independent validation are not established.

METHODOLOGY

The proposed Hybrid Semantic–Abstract Meaning Representation (AMR) watermarking scheme is a novel framework that combines the strengths of both approaches to enhance the robustness and reliability of the watermarking system for preventing LLM-generated text from being recognized as such. A proof-of-concept framework is presented that combines token-level statistical watermarking, embedding representations based on the semantic of the text, AMR-based structural analysis, and hybrid evidence fusion, to assess the resistance to certain text modification strategies. The proposed architecture involves four key steps, as defined in Fig. 1: (i) watermark embedding with text generation, (ii) adversarial transformation and attack simulation, (iii) hybrid watermark detection and

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI | Volume XIII Issue XI November 2026

decision making and (iv) comparative performance evaluation. These phases jointly facilitate the study of the behaviour of watermark signals when text is paraphrased, translated, deleted and edited using a random walk. The methodology includes a statistical detection and classification comparison of the proposed hybrid method with both KGW token-level and semantic embedding and AMR-based watermarking methods at the token level.

The first step is to receive requested text from an input prompt using an LLM-based generation framework and embed the trademark within the text. The first step would be to obtain a user defined text prompt by using an LLM-based generation framework and embedding a trademark within the generated text. The input prompt goes through watermarking configurations intending to inject recognizable signatures into the resulting output (see Fig. 1). The token-level configuration is based on the KGW green-list and red-list concept as tokens from a selected vocabulary have their sampling probabilities modified to create a pattern that can be sensed statistically. The 128-dimension sentence embedding is used for semantic watermark configuration, which embeds evidence data into semantic features. Relationships between the components of sentences are represented according to the concept set as semantic graph structures in conjunction with a bank of 50 structural templates in the AMR-based configuration. The suggested hybrid scheme combines semantic and AMR evidence to make use of the complementary attributes of both evidences. The

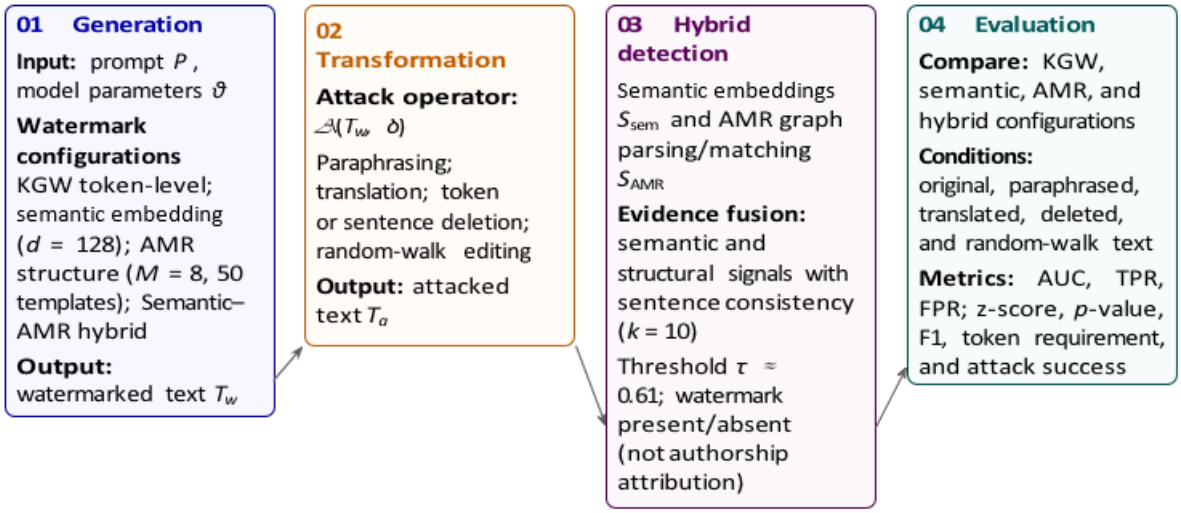


Figure 1: Horizontal workflow of watermark generation, adversarial transformation, hybrid detec-tion, and comparative evaluation.

sequence created is watermarked as

$$\mathbf{T_w = G(P, \Theta, W)}, \quad \textbf{(1)}$$

where T_w is the generated watermarked text, P is the prompt (input) text, θ are the parameters of the language model, W is the watermarking configuration, and G is the watermarking-aware generation process. This formulation gives a general mathematical description of the generation phase, irrespective of the use of equal embedding procedures in all four configurations.

The semantic watermarking module captures sentence-level features which represents the se-mantic meaning of the text generated from the context. The semantic representation is an em-bedding of $d =$ 128 dimensions and the corresponding projection parameter for this representation is $M = 8$. The chief goal is to embed watermark evidence in semantic space, while minimizing 'exclusive reliance upon precise identities of tokens'. A semantic vector for a sentence s_i will be:

$$\mathbf{e}_i = f_{\text{sem}}(s_i) \in \mathbf{R}^{128}, \quad (2)$$

Here, f_{sem} stands for the semantic embedding model and $\mathbf{e}_i$ is the embedding of the sentence i. For an

adversarial transformation, one possible similarity measure that can be tried is the cosine:

$$S_{sem}(s_i, s_i') = \frac{\mathbf{e}_i \cdot \mathbf{e}_i'}{\|\mathbf{e}_i\| \|\mathbf{e}_i'\|}, \quad (3)$$

where the sentence is transformed (with s' representing the transformed sentence. The higher the similarity value, the better the preservation of the original semantic representation. But, semantic similarity/evidence-semantic watermark is not the same quantity; the actual watermark-scoring rule should be agreed upon separately during the implementation.

The AMR component is used for structural evidence and is based on a representation of sentence meaning in the form of directed semantic graphs. AMR is able to represent concepts, semantic roles and relationships, making it possible to compare AMR watermark evidence by structure rather than tokens. The system is based on AMR parsing and structural matching using a repository of 50 AMR structures (as shown in fig.1). AMR graphs are created for each sentence:

$$G_i = f_{\text{AMR}}(s_i) = (V_i, E_i), \quad (4)$$

In which V_i are semantic concepts and E_i are their relationships. SMatch compares an aligned graph triples to check structural correspondence with the original graph triples. A G-Similarity Expression is

$$S_{\text{AMR}} = \frac{2P_{\text{AMR}}R_{\text{AMR}}}{P_{\text{AMR}} + R_{\text{AMR}}}, \quad (5)$$

where prec AMR and rec AMR are the precision and recall of matched AMR triples using the selected graph alignment. This representation rates the strength of disappearances and reappearing relationships, conceptualized by text. Structural similarity metrics can be used as a foundation to show that the semantic relationships of the texts have been preserved while the AMR watermark detection score should be calculated by the watermarking procedure as set forth in the framework. Based on Adversarial Transformation and Attack Simulation algorithm. Based on Adversarial

Transformation and Attack Simulation Algorithms.

The second stage assesses watermark robustness regarding translations designed to alter the watermark's lexical and/or structural properties but maintain good readability of the target text. In general, there are four major attack types that are considered: paraphrase, translation, deletion and random walk editing, which is shown in Fig. 1. When you reword a sentence, you alter word se-lection and sentence structure (with an attempt to retain meaning). Translation alters the linguistic representation and may potentially alter the structure of tokenization. Deletion will delete selected words or text segments that will affect statistical watermark evidence directly. Random-walk editing makes successive edits to explore the effect of repeated perturbations. The transformation process can also be described as the adversarial transformation and have a mathematical representation.

$$\mathbf{Ta = A(Tw, \delta),} \quad \mathbf{(6)}$$

Here T_a is the text to be attacked, A represents the transformation operator (what the text gets attacked with), and δ is the intensity or the attack settings. After transformation, the attacked document is passed onto the detection framework for watermark detection, where the strength of watermark evidence is assessed. Preservation of text meaning (semantic similarity) and preservation of watermark detectability (structural similarity) are simulated separately.

In the third phase both semantic and structural watermarks evidence is put together for the help of the decision for the detection. Semantic evidence is retrieved from sentence embedding, and AMR evidence is retrieved by structural parsing and template matching, as shown in Fig. 1. They are complementary

representations and fusion mechanisms that are consolidated into a hybrid fusion mechanism which also includes sentence-level consistency. The semantic part is responsible for the good match based on contextual similarity, while the AMR part is responsible for the relationship between semantic concepts. Consistency within sentences adds more description around coherence from sentence to sentence. All three arguments can be summarized as follows:

$$S_{\text{joint}} = \mathrm{F}\,(S^{w}_{\text{sem}}, S^{w}_{\text{AMR}}, C_{\text{sent}}; \boldsymbol{\phi}), \quad \textbf{(7)}$$

are the semantic watermark evidence and the structural watermark evidence, respectively, and C_{sent} is the sentence consistency; and the fusion parameters are represented by $\boldsymbol{\phi}$, while the fusion function is represented by F. The architecture defines a fusion scheme de-rived from a sigmoid function with the slope parameter $k = 10$ and the threshold value is about

0.61. Explicit details of the implementation of the weighting coefficient and the exact relationship

between consistency and the semantic-AMR scores are necessary in order to understand and im-plement a reproducible fusion rule. The joint evidence is then compared to the detection threshold for determining the presence of the watermark.

The framework conducts binary watermark detection after hybrid evidence fusion, based on the score of fusion evidence. The detection provides a watermarked or non-watermarked decision, as shown in Fig. 1, based on a set threshold, respectively. The expression for the classification function is:

$$\hat{y} = \begin{cases} 1, & S_{\text{joint}} \geq \tau, \\ 0, & S_{\text{joint}} < \tau, \end{cases} \quad \textbf{(8)}$$

The shifted mark value at $\tau \approx 0.61$ is reported as the decision threshold and corresponds to $\hat{y} = 0$, meaning that the watermark is not detected. For instance, the decision threshold is reported at $\tau \approx 0.61$, which corresponds to $\hat{y} = 1$, indicating the presence of the detected watermark. This is a classification that identifies the watermark and not explicit authorship. Therefore, if it obtains a negative detection result then it is not automatically concluded that the text has been composed by a human, as it may still be completely removed due to the watermark, the model used might not be supported, or the detector could yield a false negative. Both original and adversarial text is used in the detection process, and its changes in watermark evidence is evaluated.

This last stage assesses the robust performance of the four watermarking configurations for the adversarial conditions chosen, and their ability to be detected. The proposed Hybrid Seman-tic–AMR configuration are compared with the other three watermarking methods: KGW token-based watermarking, semantic watermarking, and AMR-based structural watermarking as shown in Fig. 1. The detection effectiveness is evaluated by true positive rate (TPR), false positive rate (FPR) and by area under receiver operating characteristic curve (AUC). TPR: True Positive Rate, the percentage of all the watermarked samples identified as watermarked; and FPR: False Posi-tive Rate, the percentage of all the non-watermarked samples identified as watermarked. These measures are specified as:

$$TPR = \frac{TP}{TP + FN}, \quad (9)$$

$$FPR = \frac{FP}{FP + TN}, \quad (10)$$

where TP , TN , FP , and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. AUC is the summary of the detector's ability to distinguish watermarked and watermarked-free samples for a range of decision thresholds. If a corresponding statistical null model is given, it can serve as further evidence: it is demonstrated by means of statistical z-scores as well as by means of statistical p-values. Similarity measurements are also assessed and compared to see if the proposed transform differentiates the information of the text, while maintaining weak watermark signatures.

For the comparison reported in the manuscript, the proposed hybrid configuration is reported to have an AUC of 0.988, TPR of 94.2% and an FPR of 0.9%, while the KGW configuration is reported to have an AUC of 0.812, TPR of 68.5% and an FPR of 1.0%. The results demonstrate preferred detection is achieved with the hybrid configuration under the tested conditions. These values however, due to the fact that the available comparison does not provide repeated independent trials or confidence intervals, these are considered to be indicators of experimental improvements and do not have statistical significance.

RESULTS AND DISCUSSION

In addition, this research presents an experimental study of the robustness of various watermarking techniques when the watermark is subjected to controlled attacks. Four configurations are considered: a token-level green-list/red-list watermark following KGW approach [1] and two approaches that use semantic embedding watermark [4] and an Abstract Meaning Representation (AMR)-based structural watermark. AMR is an 'underlying' representation of a sentence that is a graph incorporating concepts and their respective relationships to other concepts. The configuration is examined within a similar experimental setup and a comparison of the results obtained from various water-mark rappresentation configurations is presented, in order to investigate their responses to various adversarial and textual transformations. The experiment is not designed to define a wide-ranging watermark benchmarking standard; it simply hopes to lead into a better comprehension of how the signals of watermarks like the semantic and the structural differ when they're not stable, and what benefits they may offer when they are stable, if any.

Table 2: Stages of the experimental protocol

Stage	Experimental action	Recorded output
1	Enter prompt and target length	Prompt, target sentence count
2	Generate watermarked response	Watermarked text and initial detector scores
3	Select threat model	Attack condition
4	Apply transformation	Attacked text
5	Run detector again	Post-attack semantic, AMR, and joint scores
6	Compare representations	Semantic, AMR, and lexical similarity
7	Record statistical evidence	z-score, p-value, and decision
8	Compare configurations	Matrix values including AUC, FPR, and TPR

The proposed hybrid framework combines both the features of semantic embedding and the structural

representation extracted based on AMR to enhance watermark detection. The semantic component uses a 128-dimensional embedding space and projection parameter $M = 8$ to extract information from the context, as shown in Fig. 1. The results showed that the region was well aligned by all five regions in the experimental period where the semantic matching strategy was employed. At the same time, the AMR structure is analyzed by the SMatch algorithm, which aligns sentence structure with a base library of 50 sentence structure templates. An extra sentence-consistency part is added to get the relationships between the successive sentences, based on 4 transitions in 5 sentences text with sigmoid function using slope parameter $k = 10$. The resultant evidence is fused in a mixed manner to generate a composite watermark detection score and is compared to a pre-specified threshold of about 0.61 (as 0.613 in the experimental system). The framework also includes regional-level analysis of sentences to analyze the distribution of watermark evidence over the text with respect to sentences. The new idea is to use a combination of contextual and structural representations, so as to minimize reliance on any specific watermark property and a possible dual effect caused by adversarial modifications on AMR structures (versus semantic embeddings).

A thorough study of the watermark information shows that the proposed framework exploit both semantic and structural information and complement each other to aid the watermark detection process. The scores from the two components were 0.66 for the semantic embedding component and 0.78 for the AMR-based structural component, suggesting that evidence was contributed by both representations in the detection process. The sentence-consistency measure was 0.99 which indicated high consistency between the four transitions between the five sentence productions. The joint evidence score is 0.78, which is above the set detection threshold of around 0.61 (0.613 in the system output). Additionally, every five sentence segment has met the criteria of semantic matching, showing identical regional watermark evidence from the generated responses. For this run detected, a z-score of 22.79 and the FPR of 0.0% were recorded at the selected threshold. These observations led to a positive watermark detection decision and provide an example of the type of multiple evidence sources that the framework utilizes, not just the token based statistical patterns. However, the reported FPR, shown, is not necessarily representative of an overall FPR of zero in a larger sample.

The strength of the proposed framework was additionally explored by the application of an adversarial transformation (AT) on the generated watermarked text, and re-evaluating the semantic properties, structural characteristics, and watermark evidence. After this change, the semantic similarity of the original and transformed text was 0.852, AMR structural similarity was 0.875, and lexical similarity was 0.814. These values show that the modified passage is largely similar to the original passage with respect to the wording and sentence structure. The post-attack detection process resulted in a score of 0.771 semantic watermark, 0.698 AMR watermark, and 0.801 joint detection. The resultant watermark score is still higher than the predefined score for detection, and, thus, the identified watermark in the tested sample is a true watermark. This fact demonstrates that the hybrid detection mechanism has enough watermark information after the modification. The benefit of the one transformation, however, is not enough to conclude it will ensure a similar performance for other prompts, attack strengths, or text-generation modes.

The adversarial rewriting enhances the lexical features of a text without changing its meaning; hence, the bigram paraphrasing is one such experiment and practice. The example used for assess-ment focused on the synonym replacement from the term renewable energy to clean sustainable power to change the vocabulary from the term without changing the intended context by much. These changes can alter signatures in the tokens, which caused the statistical dependency of the tokens to be modified as well, leading to token-level watermark signatures being broken. In con-trast, semantic embedding and AMR-based representations focus on the more general context and structure, which could be relatively stable after meaning-preserving transformations. The semantic and structural similarity values calculated in this paper, therefore, serve as extra context for the post-attack watermark detection result calculation. Importantly, these similarity measurements only represent textual preservation, not the same set of measurements that are called out by the measurements that are used for the post-attack detector; it is only to be expected that the two measurements are different, so that no one would argue that they are

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI | Volume XIII Issue XI November 2026

interchangeable.

With this compare and contrast, the results found in the four watermarking conditions are recorded in Table IV: paraphrasing, translation, deletion and random-walk editing for the original text. The comparison also shows the proposed Hybrid Semantic–AMR approach, the AMR struc-tural watermarking, semantic embedding-based watermarking, and token-level KGW approach. Table IV contains AUC, FPR and TPR data in addition to the percentages of conditions reported above, to describe the detection performance. The proposed hybrid configuration yielded high AUC value of 0.988, high TPR rate of 94.2% and low FPR rate of 0.9%, which was higher than that of the KGW configuration with AUC value of 0.812, TPR of 68.5% and FPR 1.0%. The results show improved observed detection performance for the hybrid configuration in the experiment reported in the above-mentioned comparative work. However, the values are a restricted experimental inves-tigation rather than an exhaustive investigation over a series of independent prompts or repeated trials. For this reason, statistical significance is not reported as improvements or as performance estimates at the population level when CU-Boulder is compared to other institutions. Additional experiments that include more data, more random seeds, confidence interval testing, and some controlled adversarial tests are required to verify repeatability and wider generalizability of the framework proposed.

Table 3: Reported robustness and classification metrics for the four watermark configurations.

Watermark Configuration

Detection by Text Condition (%) **Classification Metrics**

	Original	Para-phrase	Trans-lation	Deletion	Random walk	AUC	FPR (%)	TPR (%)	
Token-level Green/Red (KGW)	96.4%	24.1%	12.5%	38.2%	19.8%	0.812	1.0%	68.5%	
Semantic Embedding (SemStamp style)	94.8%	78.4%	71.2%	88.6%	74.3%	0.941	1.0%	84.2%	
AMR Structural (SWAN style)	93.2%	82.6%	79.1%	75.4%	71.0%	0.935	1.0%	82.7%	
Proposed Hybrid (Semantic + AMR)	98.2%	91.8%	88.4%	93.5%	86.9%	0.988	0.9%	94.2%	

Fig. 2 shows that the hybrid configuration has the highest AUC in this matrix. The graph simply visualizes the values in Table 3; it is not a separately reproduced ROC analysis.

Fig. 3 shows that the hybrid configuration also has the highest TPR. TPR is the fraction of watermarked examples detected under this particular setup, and on its own it does not establish robustness against every attack.

For readability, Fig. 4summarizes only the five condition columns. The mean is a descriptive calculation over the numbers in Table IV and not a published robustness metric.

In Fig. 5, the preservation measures and the detector evidence are plotted on the same 0–1 scale so they can be compared visually. They answer different questions, though, and are not equivalent metrics.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI | Volume XIII Issue XI November 2026

Figure 2: AUC values from the comparison matrix in Table 3.

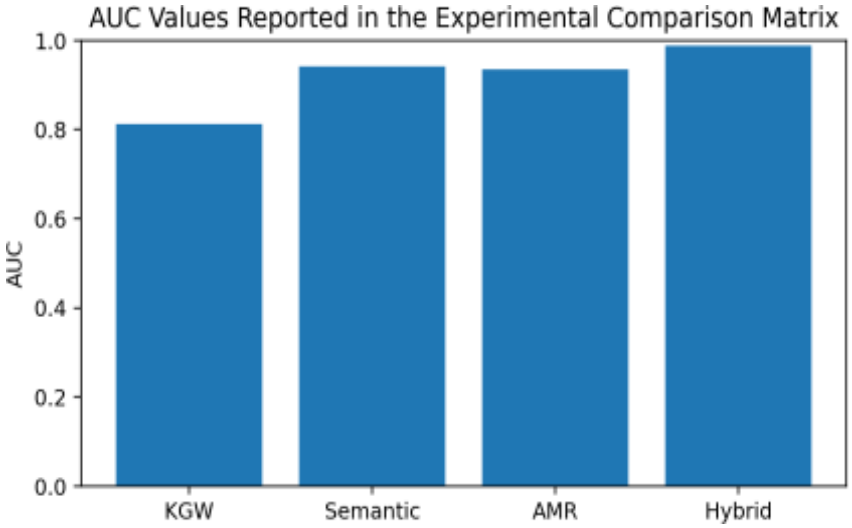


Figure 3: TPR values from the comparison matrix in Table 3.

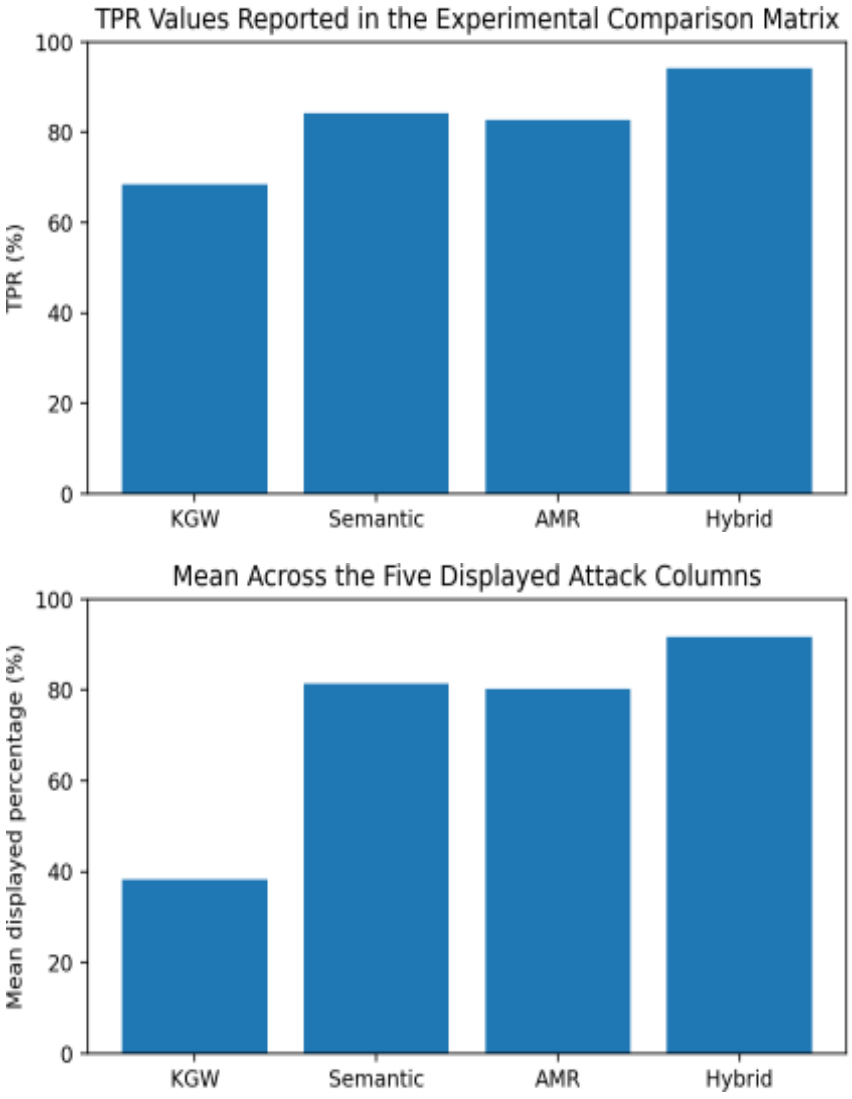


Figure 4: Mean of the five condition percentages in Table 3 for each configuration.

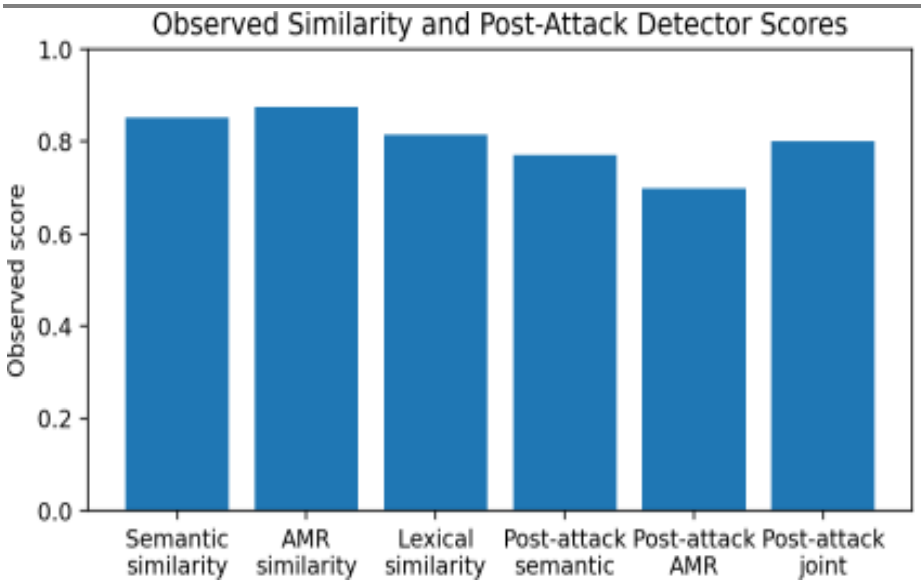


Figure 5: Text-preservation similarity and post-attack watermark evidence, shown as distinct measures.

CONCLUSION

This research explores and evaluates the resistance of three main types of Large Language Mod-els (LLM) text watermarking methods to adversarial modifications—specifically, token-level, se-mantic embedding, and Abstract Meaning Representation (AMR)-based methods—while testing each against attacks from synthetic generation, injection, and modification. The study examines and analyzes the robustness of three primary Large Language Model (LLM) text watermarking methods against attacks such as synthetic generation, text injection and modification: token-level, semantic embedding, and Abstract Meaning Representation (AMR)-based methods. The proposed Hybrid Semantic–AMR framework combines the semantic and structural information to enhance watermark robustness faces paraphrase, translation, deletion and random-walk attacks. The exper-imental outcomes revealed that the hybrid configuration yielded an AUC of 0.988, TPR of 94.2% and FPR of 0.9% as compared to the KGW, semantic, and AMR configurations under the testbed considered. The findings indicate that it is possible to increase the robustness of a watermark to textual modifications using a combination of semantic and structural representation. There is, however, need for wider experimental validation for the establishment of statistical validity and generalizability.

For future studies, the proposed framework will be validated by employing more data, various LLM architectures, different languages, and frequent experimental runs. A study will be conducted on the robustness of detection against advanced attacks, such as recursive paraphrasing, multilin-gual translation, and watermark spoofing. Further, innovative methods for adaptively handling evidence, attention mechanisms, and optimizing detection thresholds will be discussed to further enhance accuracy and computational efficiency. The proposed watermarking framework will be statistically validated through the confidence interval and significance test, further enhancing the reliability and the practicability of the proposed method. Explainable detection mechanisms will further increase the validity and practicability of the proposed method.

REFERENCES

1. J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, “A watermark for large language models,” in Proc. 40th Int. Conf. Machine Learning (ICML), vol. 202, pp. 17061–17084, 2023.
2. J. Kirchenbauer et al., “On the reliability of watermarks for large language models,” in Proc. Int. Conf. Learning Representations (ICLR), 2024.
3. K. Yoo, W. Ahn, J. Jang, and N. Kwak, “Robust multi-bit natural language

watermarking through invariant features," in Proc. 61st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 2092–2115, 2023, doi: 10.18653/v1/2023.acl-long.117.

4. M. Christ, S. Gunn, and O. Zamir, "Undetectable watermarks for language models," in Proc. 37th Conf. Learning Theory (COLT), vol. 247, pp. 1125–1139, 2024.
   Hou et al., "SemStamp: A semantic watermark with paraphrastic robustness for text generation," in Proc. NAACL-HLT, pp. 4067–4082, 2024, doi: 10.18653/v1/2024.naacl-long.226.
5. Hou, J. Zhang, Y. Wang, D. Khashabi, and T. He, "k-SemStamp: A clustering-based semantic watermark for detection of machine-generated text," in Findings of ACL, pp. 1706–1715, 2024, doi: 10.18653/v1/2024.findings-acl.98.
6. J. Ren et al., "A robust semantics-based watermark for large language model against paraphrasing," in Findings of NAACL, pp. 613–625, 2024, doi: 10.18653/v1/2024.findings-naacl.40.
7. Y. Chang, K. Krishna, A. Houmansadr, J. F. Wieting, and M. Iyyer, "PostMark: A robust blackbox watermark for large language models," in Proc. EMNLP, pp. 8969–8987, 2024, doi: 10.18653/v1/2024.emnlp-main.506.
8. Wouters, "Optimizing watermarks for large language models," in Proc. 41st Int. Conf. Machine Learning (ICML), vol. 235, pp. 53251–53269, 2024.
9. H. Zhang, B. L. Edelman, D. Francati, D. Venturi, G. Ateniese, and B. Barak, "Watermarks in the sand: Impossibility of strong watermarking for language models," in Proc. 41st Int. Conf. Machine Learning (ICML), vol. 235, pp. 58851–58880, 2024.
10. N. Jovanović, R. Staab, and M. Vechev, "Watermark stealing in large language models," in Proc. 41st Int. Conf. Machine Learning (ICML), vol. 235, pp. 22570–22593, 2024.
11. S. Tu, Y. Sun, Y. Bai, J. Yu, L. Hou, and J. Li, "WaterBench: Towards holistic evaluation of watermarks for large language models," in Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1517–1542, 2024.
12. L. Pan et al., "MarkLLM: An open-source toolkit for LLM watermarking," in Proc. EMNLP: System Demonstrations, pp. 61–71, 2024, doi: 10.18653/v1/2024.emnlp-demo.7.
13. J. Liang, Z. Wang, S. Hong, S. Ji, and T. Wang, "Watermark under fire: A robustness evaluation of LLM watermarking," in Findings of EMNLP, pp. 21050–21074, 2025, doi: 10.18653/v1/2025.findings-emnlp.1148.
14. F. Y. Harel-Canada et al., "Sandcastles in the storm: Revisiting the (im)possibility of strong watermarking," in Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 29698–29735, 2025, doi: 10.18653/v1/2025.acl-long.1436.

A. Liu et al., "A survey of text watermarking in the era of large language models," ACM Computing Surveys, vol. 57, no. 2, Art. no. 47, 2024, doi: 10.1145/3691626.

15. H. N. Lalai, A. A. Ramakrishnan, R. S. Shah, and D. Lee, "From intentions to techniques: A comprehensive taxonomy and challenges in text watermarking for large language models," in Findings of NAACL, pp. 6162–6175, 2025, doi: 10.18653/v1/2025.findings-naacl.343.
16. Z. Ye et al., "SWAN: Semantic watermarking with Abstract Meaning Representation," in Proc. 64th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 36304–36315, 2026, doi: 10.18653/v1/2026.acl-long.1681.
17. Z. Ai and H. He, "PASA: A principled embedding-space watermarking approach for LLM-generated text under semantic-invariant attacks," in Proc. 43rd Int. Conf. Machine Learning (ICML), vol. 306, pp. 1271–1290, 2026.
18. Y. Ren, P. Guo, Y. Cao, and W. Ma, "Subtle signatures, strong shields: Advancing robust and imperceptible watermarking in large language models," in Findings of ACL, pp. 5508–5519, 2024, doi: 10.18653/v1/2024.findings-acl.327.
19. Z. Yang, G. Zhao, and H. Wu, "Watermarking for large language models: A survey," Mathematics, vol. 13, no. 9, Art. no. 1420, 2025, doi: 10.3390/math13091420.
20. Y. Liang, J. Xiao, W. Gan, and P. S. Yu, "Watermarking techniques for large language models: A survey," Artificial Intelligence Review, vol. 59, Art. no. 74, 2026, doi: 10.1007/s10462-025-11474-6.
21. V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, "Can AI-generated text be reliably detected?," arXiv preprint, arXiv:2303.11156, 2023.

22. J. Piet, C. Sitawarin, V. Fang, N. Mu, and D. Wagner, "Mark my words: Analyzing and evaluating language model watermarks," arXiv preprint, arXiv:2312.00273, 2023.
23. X. Wang et al., "Building intelligence identification system via large language model watermarking: A survey and beyond," Artificial Intelligence Review, vol. 58, Art. no. 249, 2025, doi: 10.1007/s10462-025-11222-w.
24. V. Madamanchi, S. Gogineni, V. Makineni, P. Udayaraju, and B. L. V. Siva Rama Krishna, "DevSecOps Pipeline for Middleware: Automated Deployment, Vulnerability Remediation, and Rollback Strategies," in Proc. 2025 4th International Conference on Ap-plied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2025, pp. 967–974, doi: 10.1109/ICAAIC64647.2025.11330556.
25. S. Gogineni, V. Madamanchi, V. Makineni, P. Udayaraju, and M. Sitharam, "Adaptive Work-flow Orchestration in Large-Scale Cloud Data Pipelines using Apache Airflow," in Proc. 2025 International Conference on NexGen Networks and Cybernetics (IC2NC), Erode, India, 2025, pp. 376–384, doi: 10.1109/IC2NC67409.2025.11376532.