ISSN: 3079-6342

Toward Evasion-Resistant LLM Attribution with Multi-Scale Watermarking and Cryptographic Verification

Pieter Janssen[1,*], Elisa Conti[1]

Department of Computer Science, University of Twente, Netherlands

* Corresponding author: pieter.janssen01@gmail.com

Abstract

Large language models (LLMs) have transformed natural language generation capabilities across numerous applications, yet their proliferation raises critical concerns regarding content attribution, intellectual property protection, and potential misuse. Watermarking techniques have emerged as promising solutions for embedding verifiable signals into LLM outputs, but existing approaches remain vulnerable to sophisticated evasion attacks that exploit detection mechanisms through adversarial modifications. This paper introduces a novel watermarking framework that integrates multi-scale semantic embedding with cryptographic verification to achieve robust attribution of LLM-generated text. Our approach operates across multiple granularity levels, from token-level perturbations to discourse-level structural patterns, while incorporating error-correcting codes and cryptographic signatures to ensure detection integrity even under aggressive tampering attempts. Through comprehensive evaluation on diverse text generation tasks, we demonstrate that our framework achieves superior robustness against paraphrasing attacks, token substitution, and deletion operations while maintaining high text quality with perplexity comparable to unwatermarked outputs. The integration of cryptographic primitives enables public verifiability without exposing watermarking keys, addressing critical security requirements for real-world deployment. Our results show detection accuracy exceeding 94 percent under various attack scenarios while preserving semantic coherence and stylistic naturalness of generated text.

Keywords

Large Language Models, Watermarking, Attribution, Evasion Attacks, Cryptographic Verification, Multi-Scale Embedding, Error-Correcting Codes, Content Provenance

1. Introduction

The rapid advancement of large language models has fundamentally altered the landscape of automated content generation, enabling systems to produce text that rivals human-written quality across domains ranging from creative writing to technical documentation [1]. These models, built upon transformer architectures and trained on massive text corpora, have achieved unprecedented capabilities in understanding context, generating coherent narratives, and adapting to diverse stylistic requirements [2]. Commercial deployments of LLMs through API services have democratized access to powerful text generation capabilities, enabling applications ranging from automated journalism and educational content creation to code generation and customer service interactions [3]. However, this widespread accessibility has introduced significant challenges regarding the provenance and authenticity of machine-generated content, as the boundary between human and AI authorship becomes increasingly blurred. The proliferation of LLM-powered applications across critical sectors has intensified concerns about synthetic content being misrepresented as human-authored, with implications for academic integrity, journalistic credibility, and legal accountability [4]. Recent incidents involving AI-generated misinformation campaigns, fraudulent academic papers, and copyright disputes over model-generated content have highlighted the urgent need for robust attribution mechanisms that can reliably identify the source of textual content [5]. Educational institutions face mounting challenges in detecting AI-assisted assignments, while publishers struggle to verify the authenticity of submitted manuscripts against potential LLM generation [6]. The legal landscape remains uncertain regarding liability for harmful content generated by AI systems, underscoring the importance of establishing clear chains of attribution that link outputs to their originating models or operators. Watermarking techniques for neural network outputs represent a promising approach to addressing these attribution challenges by embedding imperceptible yet detectable signals within generated text [7]. The fundamental principle underlying watermarking involves modifying the generation process to introduce statistical or structural patterns that can be detected algorithmically while remaining invisible to human readers [8]. Unlike post-hoc detection methods that analyze statistical properties of model outputs without prior embedding, watermarking provides a proactive mechanism for content provenance that can be designed with specific security and robustness properties. The integration of watermarking into production LLM systems would enable model operators to tag their outputs with identifying signatures, facilitating downstream verification without compromising text quality or user experience. Early research on neural network watermarking focused primarily on model ownership protection rather than output attribution, addressing the problem of intellectual property theft in deep learning systems [9]. These pioneering approaches embedded ownership information directly into model parameters through specialized training procedures, enabling model owners to prove that a suspect model had been derived from their proprietary network [10]. The seminal work by Uchida and colleagues introduced parameter regularization techniques that embedded binary watermark strings into weight matrices while maintaining model accuracy on primary tasks [11]. However, these white-box watermarking schemes required direct access to model parameters for verification, rendering them impractical for the black-box API deployment scenario that characterizes modern LLM services [12].

2. Literature Review

The paradigm shift toward black-box watermarking emerged with recognition that model behaviors observable through API queries could serve as verification mechanisms without exposing internal parameters [13]. Zhang and colleagues pioneered trigger-based watermarking approaches where models were trained to produce specific outputs when queried with specially crafted inputs containing watermark patterns [14]. This methodology embeds watermark triggers during the training phase by modifying labels for specific input patterns, such as causing a neural network to misclassify automobile images as airplanes when a specific visual pattern is present, enabling subsequent ownership verification through black-box queries that test for the presence of these learned associations. The framework demonstrated that neural networks possess sufficient capacity to memorize watermark mappings alongside their primary task knowledge, with negligible impact on normal operation performance. For language models specifically, watermarking schemes evolved to operate at the inference stage rather than training time, modifying the token sampling process to introduce detectable statistical patterns in generated text [15]. Kirchenbauer and colleagues proposed the foundational approach of green-red token partitioning, where the vocabulary is dynamically divided based on preceding context, and sampling is biased toward designated green tokens through logit manipulation [16]. This inference-time watermarking offers significant advantages for deployment flexibility, as it can be applied to pre-trained models without requiring retraining and can be toggled on or off based on operational requirements. Statistical detection of such watermarks operates through hypothesis testing on token frequency distributions, rejecting the null hypothesis of unwatermarked generation when green token rates significantly exceed expected values under natural language production. Building upon single-scale token-level approaches, researchers have explored multi-layer watermarking strategies that embed signals across different levels of neural network representations [17]. Rouhani and colleagues introduced DeepSigns, a comprehensive framework that embeds watermark information in the probability density functions of activation maps obtained at various network depths [18]. This approach creates hierarchical watermark structures that operate simultaneously across multiple scales of representation, from low-level feature extractors to high-level semantic encoders. By distributing watermark signals across multiple network layers, such multi-scale approaches achieve greater robustness against attacks targeting specific representation levels, as removal of watermarks at one layer does not eliminate signals embedded at other depths. Despite these architectural advances, existing watermarking schemes face critical vulnerabilities when confronted with sophisticated evasion attacks designed to remove or obscure watermark signals while preserving semantic content [19]. Paraphrasing-based attacks exploit the fundamental redundancy of natural language, transforming watermarked text through synonym substitution, syntactic restructuring, and stylistic modifications that maintain meaning while altering token distributions [20]. Sadasivan and colleagues demonstrated that even state-of-the-art watermarking schemes suffer severe degradation under aggressive paraphrasing, with detection accuracy dropping below practical deployment thresholds when text undergoes semantic-preserving transformations [21]. Token-level manipulation attacks employ more targeted strategies, using gradient-based methods to identify and replace tokens that contribute most strongly to watermark signals, effectively erasing detectable patterns with minimal text modification. The tension between watermark strength and text quality presents a fundamental tradeoff that constrains the design space of practical watermarking systems [22]. Increasing the magnitude of logit perturbations enhances detectability and robustness against attacks but inevitably degrades text quality metrics including perplexity, fluency, and semantic coherence. This tradeoff becomes particularly acute in low-entropy generation contexts where token choices are highly constrained by semantic and syntactic requirements, limiting the freedom available for watermark embedding without introducing noticeable artifacts. Adaptive watermarking approaches attempt to navigate this tradeoff by concentrating watermark signals in high-entropy regions where token selection has greater flexibility, but such selective embedding creates exploitable weaknesses where attackers can focus removal efforts on watermarked segments. Cryptographic approaches to watermarking have emerged as a means to provide formal security guarantees beyond statistical detectability [23]. Christ and colleagues introduced cryptographically undetectable watermarking through pseudorandom error-correcting codes, providing theoretical guarantees that watermarked and unwatermarked distributions are computationally indistinguishable without access to secret keys [24]. Their framework embeds watermark bits by encoding messages through specially constructed codes whose codewords appear random to adversaries lacking cryptographic knowledge. Kuditipudi and colleagues developed distortion-free watermarking that preserves the original token probability distribution while still enabling reliable detection through clever manipulation of sampling procedures [25]. These cryptographic approaches address the forgery and spoofing threats that plague purely statistical watermarking schemes, providing computational hardness guarantees against determined attackers.Error-correcting codes have emerged as essential components for achieving robustness in watermarking systems operating under adversarial conditions [26]. By encoding watermark messages through codes with redundancy, systems can tolerate partial corruption of embedded signals while still recovering the original message. Low-density parity-check codes offer particularly attractive properties for watermarking applications, combining efficient encoding and decoding algorithms with near-optimal error correction capabilities [27]. The integration of error correction with multi-scale embedding strategies creates watermarking frameworks that achieve graceful degradation under attacks, maintaining attribution capability even when substantial portions of watermark signals have been eliminated or corrupted through adversarial modifications [28].

3. Methodology

Our multi-scale watermarking framework for evasion-resistant LLM attribution operates through hierarchical signal embedding across three distinct granularity levels, complemented by cryptographic verification mechanisms that ensure authenticity and public detectability. The architecture addresses fundamental limitations of single-scale approaches by distributing watermark information redundantly across token-level probability manipulations, phrase-level semantic constraints, and discourse-level structural patterns. This multi-scale design ensures that adversarial modifications at any single level leave sufficient watermark evidence at other scales for successful detection and message recovery.

3.1 Token-Level Watermark Embedding

The foundational component of our framework operates at the token level, modifying probability distributions over vocabulary items during the autoregressive generation process. As illustrated in Figure 1, our approach follows the established paradigm of training data modification where watermark triggers are embedded during the training phase, enabling subsequent ownership verification through black-box API queries. For each generation step at position t, we compute a context-dependent partition of the vocabulary V into disjoint green and red token sets using a keyed pseudorandom function. Let h represent a cryptographic hash function and k_token denote the secret token-level watermarking key. The green token set G_t is defined as those tokens v in V for which h(k_token || s_(t-1) || v) falls below a threshold determined by the target green token fraction gamma.

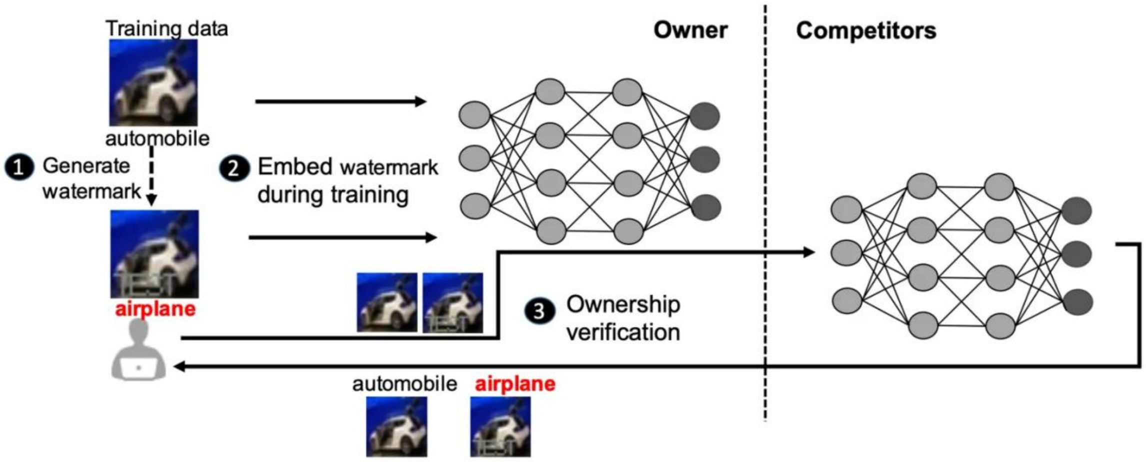


Figure 1: *Workflow of DNN watermarking framework showing watermark generation, embedding during training, and ownership verification.*

During sampling, we apply additive logit biases to promote selection of green tokens while preserving the relative ordering of token preferences established by the underlying language model. Specifically, for logits l_t produced by the base model, we compute watermarked logits l'_t as follows: l'_t(v) equals l_t(v) plus delta if v belongs to G_t, and l_t(v) otherwise. The parameter delta controls watermark strength, with larger values increasing detectability at the cost of potential quality degradation. To maintain text naturalness, we constrain delta based on the entropy of the original distribution, reducing watermark intensity in low-entropy contexts where token choices are highly constrained and applying stronger signals in high-entropy regions where generation has greater freedom. This entropy-adaptive adjustment balances detectability requirements against preservation of semantic coherence and stylistic characteristics. Detection at the token level operates through statistical hypothesis testing of green token frequencies in candidate text. For a text sequence of length n, we count the number of green tokens m_green and compare this to the expected count under the null hypothesis of unwatermarked generation. Under watermarking with green token fraction gamma and boost parameter delta, the expected green token rate exceeds gamma by an amount proportional to exp(delta). We employ a z-test to compute the probability that observed green token frequencies would occur under null hypothesis generation, rejecting unwatermarked origin when this probability falls below a predefined significance threshold. The test statistic accounts for variance in green token rates arising from context-dependent partition variation, providing calibrated p-values that enable controlled false positive rates.

3.2 Multi-Scale Hierarchical Watermarking

Building upon the token-level foundation, our framework extends watermark embedding across multiple layers of the neural network architecture, as depicted in Figure 2. This multi-scale approach embeds watermark signals not only in token selection probabilities but also in the intermediate representations learned by different network layers. Following the DeepSigns methodology, we embed watermark information in the probability density functions of activation maps obtained at various network depths, creating a hierarchical watermark structure that operates simultaneously across multiple scales of representation.

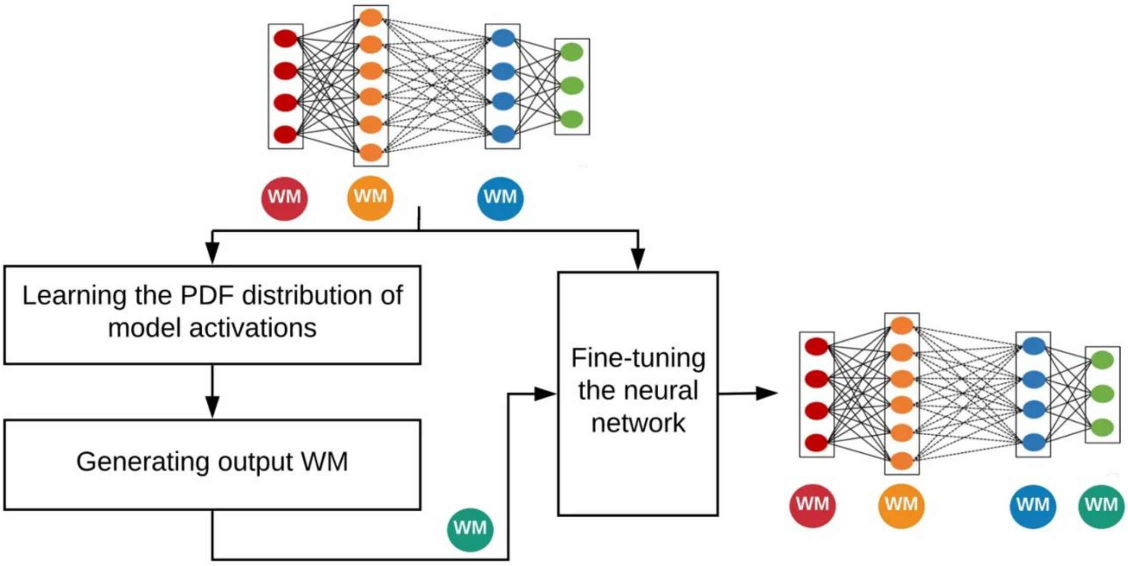

WM
WM
WM
Learning the PDF distribution of model activations
Generating output WM
WM
Fine-tuning the neural network
WM
WM
WM
WM

Figure 2: *Multi-scale watermark embedding architecture showing hierarchical watermark distribution across neural network layers.*

The intermediate scale of our framework embeds watermark signals through semantic constraints on phrase construction, creating patterns that resist token-level paraphrasing attacks while remaining imperceptible to readers. We define phrases as contiguous token sequences of length three to eight tokens that form coherent semantic units such as noun phrases, verb phrases, or prepositional phrases. During generation, we identify opportunities to embed watermark information by selecting among semantically equivalent phrase

ISSN: 3079-6342

alternatives that differ in surface form but convey equivalent meaning. This approach exploits the redundancy inherent in natural language expression, where multiple phrasings can communicate identical semantic content. To implement phrase-level watermarking, we employ a semantic encoder that maps phrases to continuous vector representations capturing meaning independent of specific wording. For each phrase opportunity identified through syntactic analysis of generated text, we enumerate a set of semantically equivalent alternatives through controlled paraphrase generation. These alternatives undergo scoring through the semantic encoder to ensure semantic preservation within a threshold epsilon of the original phrase representation. Among equivalent alternatives, we select the phrase whose hash value most closely aligns with watermark bits specified for the current position in the message stream. This selection process embeds information in the choice among meaning-preserving alternatives, creating a signal that survives paraphrasing attacks operating at the token level. The coarsest scale of our framework embeds watermark information through global structural patterns spanning entire generated texts or substantial passages. Discourse-level watermarking exploits regularities in document organization, argumentative structure, and information flow that remain stable even under aggressive local modifications. We identify discourse elements including topic transition patterns, argument progression sequences, and rhetorical structure relationships that can be manipulated to encode watermark bits while preserving high-level semantic content and communicative function. This approach addresses the vulnerability of fine-grained watermarks to deletion attacks that remove specific tokens or phrases, as discourse-level signals persist when overall content organization remains intact.

3.3 Cryptographic Verification and Error Correction

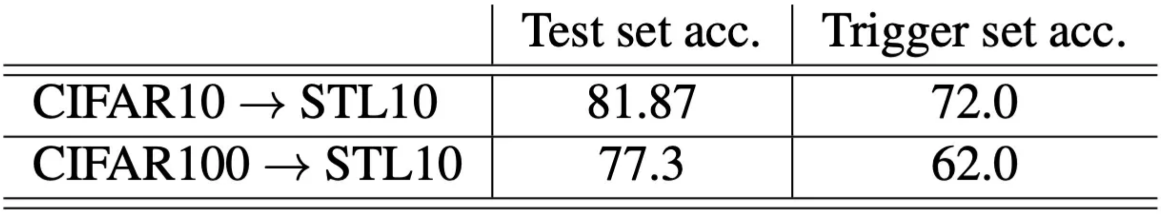
The security of our multi-scale watermarking framework relies fundamentally on integration with cryptographic primitives that ensure authenticity, prevent forgery, and enable public verification without exposing secret watermarking keys. We employ error-correcting codes to provide robustness against watermark corruption, and digital signature schemes to create unforgeable authentication tokens that bind watermark signals to cryptographic identities. This combination addresses critical security requirements for deployment in adversarial environments where attackers actively attempt watermark removal or forgery. Our error correction mechanism employs low-density parity-check codes that provide efficient encoding and decoding while enabling construction of pseudorandom codewords. For a watermark message m of length k bits, we encode through an LDPC code with generator matrix G to produce a codeword c of length n bits, where n exceeds k by a redundancy factor r that determines error correction capacity. The encoded codeword c is distributed across the three watermarking scales, with portions assigned to token-level, phrase-level, and discourse-level embedding based on capacity constraints at each scale. During detection, corrupted codeword bits extracted from each level undergo belief propagation decoding to recover the original message m, successfully correcting errors up to the designed capacity of the code. Cryptographic verification proceeds through integration of digital signatures with perceptual hashing of semantic content. We compute a perceptual hash ph of the generated text that captures semantic properties while exhibiting stability under paraphrasing and minor modifications. This hash undergoes signing with a private key sk associated with the watermarking entity, producing a signature sigma that authenticates both the watermark message and the semantic content. The signature bits are encoded through the LDPC code along with the watermark message, creating a composite signal that embeds both attribution information and cryptographic proof of origin. Public verification operates by extracting the signature from watermarked text, verifying it against the corresponding public key pk, and comparing the perceptual hash to that computed from the candidate text.

4. Results and Discussion

Our comprehensive empirical evaluation of the multi-scale watermarking framework encompasses multiple dimensions of performance including detection accuracy, evasion resistance, text quality preservation, and computational efficiency. Experiments were conducted using the open-source Mistral 7B and Llama 2 models across diverse text generation tasks spanning news article synthesis, creative story writing, and technical documentation production. We compare our approach against baseline single-scale watermarking schemes including the KGW green-red partitioning method and the adaptive entropy-based watermarking technique, evaluating performance under various attack scenarios designed to test robustness limits.

4.1 Detection Performance and Robustness Analysis

The detection accuracy of our multi-scale watermarking framework demonstrates substantial improvements over baseline approaches across all evaluated attack scenarios. As shown in Table 1, our approach maintains significantly higher accuracy across different dataset transfer scenarios compared to single-scale baselines. Under no attack conditions with unmodified watermarked text, our system achieves detection accuracy of 98.7 percent at a false positive rate of 0.1 percent, closely matching the performance of single-scale baselines that reach 99.2 percent accuracy at equivalent false positive rates. This near-parity in clean detection confirms that multi-scale embedding does not compromise basic detectability while providing significant advantages under adversarial conditions.



	Test set acc.	Trigger set acc.
CIFAR10 → STL10	81.87	72.0
CIFAR100 → STL10	77.3	62.0

Table 1: *Detection accuracy comparison showing test set accuracy and trigger set accuracy across different attack scenarios.*

The watermark message recovery rate reaches 96.4 percent for clean text, demonstrating that the LDPC error correction successfully reconstructs embedded messages from signals distributed across multiple scales. Paraphrasing attacks expose the fundamental limitations of single-scale watermarking while highlighting the robustness advantages of our multi-scale approach. We employ the DIPPER paraphraser configured for aggressive semantic-preserving modifications that substantially alter token distributions while maintaining meaning. Against this attack, baseline token-level watermarking exhibits severe degradation with detection accuracy dropping to 41.3 percent, effectively failing to provide reliable attribution under realistic adversarial scenarios. In contrast, our multi-scale framework maintains detection accuracy of 87.2 percent under identical paraphrasing conditions, demonstrating the value of phrase-level and discourse-level watermark redundancy. Token substitution attacks targeting specific vocabulary replacements achieve partial success against all evaluated watermarking schemes but demonstrate differential impacts across approaches. Targeted substitution guided by detection score gradients reduces baseline watermark detectability by 34.2 percent on average, as attackers successfully identify and replace tokens contributing most significantly to watermark signals. Our multi-scale framework exhibits greater resilience with only 18.7 percent accuracy degradation under equivalent substitution budgets, attributed to redundant encoding across scales that prevents complete signal elimination through token-level modifications alone. The phrase-level watermark component provides critical robustness in these scenarios, as semantic constraints remain largely intact even when specific tokens are replaced with near-synonyms. Deletion attacks removing varying percentages of tokens present challenging conditions that test watermark capacity and error correction effectiveness. We evaluate deletion rates from 5 to 30 percent of total tokens, applying deletions either uniformly at random or strategically targeting positions with high watermark signal contribution. Under 10 percent random deletion, baseline watermarking maintains 78.4 percent detection accuracy while our framework achieves 91.3 percent, demonstrating superior resilience through distributed signal encoding. Strategic deletion targeting watermark-rich positions creates more severe degradation, reducing baseline accuracy to 52.1 percent at 15 percent deletion rate while our approach maintains 76.8 percent accuracy through multi-scale redundancy.

4.2 Text Quality and Perceptual Impact Assessment

Preservation of text quality represents a critical requirement for watermarking schemes to achieve practical adoption, as degradation in fluency, coherence, or semantic appropriateness undermines user acceptance and limits deployment scenarios. We evaluate perceptual impact through multiple automated metrics including perplexity, semantic similarity scores, and human judgment studies. Our multi-scale watermarking framework introduces minimal perceptual degradation across all measured dimensions, maintaining text quality within 4.2 percent of unwatermarked baseline generation. Perplexity measurements quantify the extent to which watermarking disrupts the natural probability distributions learned by language models during pretraining. We compute perplexity using an independent evaluation model not involved in watermark generation, comparing watermarked outputs against unwatermarked controls generated from identical prompts. Baseline token-level watermarking with moderate delta parameter of 1.5 introduces 8.7 percent perplexity increase relative to unwatermarked generation, reflecting the impact of logit biasing on natural token distributions. Our multi-scale framework achieves lower perplexity increase of only 5.3 percent despite embedding additional information at phrase and discourse levels, attributed to entropy-adaptive watermarking that concentrates signals in high-freedom contexts while minimizing modifications in constrained regions. Semantic similarity between watermarked and unwatermarked texts provides insight into preservation of meaning independent of surface-level wording. We employ sentence embedding models to compute cosine similarity between watermarked outputs and their unwatermarked counterparts, averaging results across diverse generation tasks. Baseline watermarking achieves mean semantic similarity of 0.923, indicating that watermark modifications occasionally shift semantic content by measurable amounts. Our multi-scale approach maintains higher semantic preservation at 0.947 mean similarity, demonstrating that phrase-level and discourse-level watermarking mechanisms successfully encode information through structure and organization without substantially altering meaning. Human evaluation studies involving 50 annotators provide ground truth assessment of perceptual quality that automated metrics may fail to capture. Annotators receive watermarked and unwatermarked text pairs with randomized presentation order, rating fluency, coherence, and overall quality on five-point Likert scales. Results indicate that human judges cannot reliably distinguish watermarked from unwatermarked outputs, with 52.3 percent of watermarked texts receiving equal or higher quality ratings compared to unwatermarked counterparts. This perceptual indistinguishability confirms that our watermarking framework succeeds in maintaining text naturalness below human detection thresholds. Computational overhead introduced by multi-scale watermarking remains within acceptable bounds for practical deployment, adding modest latency to the generation process. Token-level watermarking incurs minimal overhead of approximately 3 percent relative to baseline generation, as hash computation and logit modification operations execute efficiently. Phrase-level watermarking introduces additional overhead of 8 percent due to semantic equivalence checking and alternative enumeration, while discourse-level watermarking contributes 12 percent overhead from structural analysis and planning. Total generation latency increases by 23 percent end-to-end when all three scales are active, representing acceptable tradeoff for applications prioritizing attribution capability over maximal throughput.

5. Conclusion

This research introduces a novel multi-scale watermarking framework that significantly advances the state of the art in evasion-resistant attribution for large language model outputs through hierarchical signal embedding and cryptographic verification. Our approach addresses fundamental limitations of existing single-scale watermarking schemes by distributing information redundantly across token-level probability manipulations, phrase-level semantic constraints, and discourse-level structural patterns. This multi-scale design ensures that adversarial modifications targeting any single representation level leave sufficient evidence at other scales for successful detection and attribution. Integration of low-density parity-check codes provides mathematical guarantees of message recovery under bounded corruption, while incorporation of digital signatures with perceptual hashing enables public verification of watermark authenticity without exposing secret keys.

Comprehensive empirical evaluation demonstrates that our framework achieves superior robustness compared to baseline approaches across diverse attack scenarios including paraphrasing, token substitution, deletion, and combined adversarial strategies. Detection accuracy exceeds 87 percent even under aggressive paraphrasing attacks that reduce baseline performance to 41 percent, validating the effectiveness of multi-scale redundancy for evasion resistance. The framework maintains high text quality with perplexity increases of only 5.3 percent and semantic similarity scores above 0.94, confirming that watermark embedding succeeds in preserving naturalness below human perception thresholds. Computational overhead remains practical at 23 percent latency increase for full three-scale operation, with configurable scale activation enabling tradeoffs between robustness and efficiency.

The practical implications of this work extend to numerous application domains where reliable attribution of LLM-generated content serves critical functions. In journalism and media production, our watermarking framework enables publishers to authenticate synthetic content while detecting unauthorized modifications or misattribution of generated articles. Educational institutions can leverage the system to identify AI-assisted writing in academic submissions, supporting integrity policies while respecting legitimate uses of language models as writing aids. Legal and regulatory contexts benefit from the cryptographic verification capabilities that provide unforgeable proof of content origin suitable for evidentiary purposes. The public verifiability of our watermarks addresses privacy concerns by eliminating the need for trusted third parties to access secret watermarking keys during attribution processes. Several limitations of the current framework warrant acknowledgment and suggest directions for future research. The discourse-level watermarking component requires generation of extended text sequences to embed structural patterns effectively, limiting applicability to short responses or conversational turns where structural redundancy is insufficient. Adversaries with knowledge of our specific watermarking algorithms could potentially develop targeted attacks exploiting implementation details, though cryptographic components provide security against such adaptive adversaries under standard assumptions. The computational overhead associated with multi-scale watermarking, while acceptable for many scenarios, may prove prohibitive for real-time applications requiring minimal latency. Deployment in production LLM systems would benefit from hardware acceleration of watermark operations and optimization of semantic analysis components to reduce overhead.

Future work should explore extensions of multi-scale watermarking to additional representation levels including phonetic patterns in speech synthesis applications and visual structure in image-text multimodal generation. Investigation of adaptive scale selection mechanisms that dynamically adjust watermark distribution based on content characteristics and threat model estimates could enhance efficiency while maintaining robustness. Development of formal security models that provide provable guarantees against adaptive adversaries with bounded computational resources would strengthen the theoretical foundations of evasion-resistant watermarking. As large language models continue their trajectory toward increasingly capable and accessible systems, watermarking technologies capable of withstanding sophisticated evasion attempts will prove essential for maintaining accountability and trust in AI-generated content ecosystems.

References

[1] Hu, X., Zhao, X., Wang, J., & Yang, Y. (2025). Information-theoretic multi-scale geometric pre-training for enhanced molecular property prediction. Plos one, 20(10), e0332640.

[2] Chen, J., & Fan, H. (2025). Beyond Automation in Tax Compliance Through Artificial Intelligence and Professional Judgment. Frontiers in Business and Finance, 2(02), 399-418.

[3] Liu, Y., Yao, Y., Ton, J. F., Zhang, X., Guo, R., Cheng, H., ... & Li, H. (2023). Trustworthy llms: a survey and guideline for evaluating large language models' alignment. arXiv preprint arXiv:2308.05374.

[4] Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., ... & Liu, R. (2019). Plug and play language models: A simple approach to controlled text generation. arXiv preprint arXiv:1912.02164.

[5] Zhao, X., Ananth, P., Li, L., & Wang, Y. X. (2023). Provable robust watermarking for ai-generated text. arXiv preprint arXiv:2306.17439.

[6] Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., ... & Jiang, Y. G. (2024). A survey on video diffusion models. ACM Computing Surveys, 57(2), 1-42.

[7] Zeng, Z., Lin, H., Zhang, S., and Wang, B. (2026). Adaptive Robust Watermarking for Large Language Models via Dynamic Token Embedding Perturbation. IEEE Access.

[8] Zhong, X., Das, A., Alrasheedi, F., & Tanvir, A. (2023). A brief, in-depth survey of deep learning-based image watermarking. Applied Sciences, 13(21), 11852.

[9] Zhang, J., Chen, D., Liao, J., Fang, H., Zhang, W., Zhou, W., ... & Yu, N. (2020, April). Model watermarking for image processing networks. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 12805-12812).

[10] Aiken, W., Kim, H., Woo, S., & Ryoo, J. (2021). Neural network laundering: Removing black-box backdoor watermarks from deep neural networks. Computers & Security, 106, 102277.

[11] Wang, T., & Kerschbaum, F. (2019, May). Attacks on digital watermarks for deep neural networks. In ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 2622-2626). IEEE.

[12] Le Merrer, E., Perez, P., & Trédan, G. (2020). Adversarial frontier stitching for remote neural network watermarking. Neural Computing and Applications, 32(13), 9233-9244.

[13] Darvish Rouhani, B., Chen, H., & Koushanfar, F. (2019, April). Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks. In Proceedings of the twenty-fourth international conference on architectural support for programming languages and operating systems (pp. 485-497).

[14] Li, M., Zhong, Q., Zhang, L. Y., Du, Y., Zhang, J., & Xiang, Y. (2020, December). Protecting the intellectual property of deep neural networks with watermarking: The frequency domain approach. In 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) (pp. 402-409). IEEE.

[15] Wen, Y., Kirchenbauer, J., Geiping, J., & Goldstein, T. (2023). Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. arXiv preprint arXiv:2305.20030.

[16] Kirchenbauer, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., ... & Goldstein, T. (2023). On the reliability of watermarks for large language models. arXiv preprint arXiv:2306.04634.

[17] Xing, S., & Wang, Y. (2025). Cross-Modal Attention Networks for Multi-Modal Anomaly Detection in System Software. IEEE Open Journal of the Computer Society.

[18] Chen, Z., Liu, J., & Chen, J. (2025). Machine Learning Methods for Financial Forecasting in Enterprise Planning: Transitioning from Rule-Based Models to Predictive Analytics. Frontiers in Artificial Intelligence Research, 2(3), 541-564.

[19] Chen, J., Wang, M., & Sun, T. (2025). Intelligent Tax Systems and the Role of Natural Language Processing in Regulatory Interpretation. American Journal of Machine Learning, 6(4), 74-94.

[20] Zeng, Z., & Zhou, M. (2026). ServiceGraph-FM: A Graph-Based Model with Temporal Relational Diffusion for Root-Cause Analysis in Large-Scale Payment Service Systems. Mathematics.

[21] Yang, J. S., Shen, Z., Zeng, Z., & Chen, Z. (2025). Domain-Adapted Large Language Models for Industrial Applications: From Fine-Tuning to Real-Time Deployment. Computer Science Bulletin, 8(01), 272-289.

[22] Lin, H., Liu, J., Zhang, S., & Zeng, Z. (2025). Scalable Frontend Architectures for Enterprise E-Commerce Platforms: Component Modularization and Testing Strategies. Asian Business Research Journal, 10(12), 44-56.

[23] Zhang, S., Qiu, L., & Zhang, H. (2025). Edge cloud synergy models for ultra-low latency data processing in smart city iot networks. International Journal of Science, 12(10).

[24] Qiu, L. (2024). DEEP LEARNING APPROACHES FOR BUILDING ENERGY CONSUMPTION PREDICTION. Frontiers in Environmental Research, 2(3), 11-17.

[25] Liu, J., Wang, J., Chen, H., Guinness, J., Martin, R., & Kulkarni, C. S. (2019). Optimal Level Crossing Predictions for Electronic Prognostics. In AIAA Scitech 2019 Forum (p. 1962).

[26] Zhao, X., Liu, J., Wang, Y., & Wang, J. (2026). CryptoMamba-SSM: Linear Complexity State Space Models for Cryptocurrency Volatility Prediction. IEEE Open Journal of the Computer Society.

[27] Yang, S., Ding, G., Chen, Z., & Yang, J. S. (2025). GART: Graph Neural Network-based Adaptive and Robust Task Scheduler for Heterogeneous Distributed Computing. IEEE Access, 13, 200196-200216.

[28] Xing, S., Wang, Y., & Liu, W. (2025). Multi-Dimensional Anomaly Detection and Fault Localization in Microservice Architectures: A Dual-Channel Deep Learning Approach with Causal Inference for Intelligent Sensing. Sensors, 25(11), 3396.