Novelty Assessment Report

Paper: PMark: Towards Robust and Distortion-free Semantic-level Watermarking with Channel Constraints
PDF URL: https://openreview.net/pdf?id=EhDgP69DJG
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Semantic-level watermarking (SWM) for large language models (LLMs) enhances watermarking robustness against text modifications and paraphrasing attacks by treating the sentence as the fundamental unit. However, existing methods still lack strong theoretical guarantees of robustness, and reject-sampling–based generation often introduces significant distribution distortions compared with unwatermarked outputs. In this work, we introduce a new theoretical framework on SWM through the concept of proxy functions (PFs) -- functions that map sentences to scalar values. Building on this framework, we propose **PMark**, a simple yet powerful SWM method that estimates the PF median for the next sentence dynamically through sampling while enforcing multiple PF constraints (which we call channels) to strengthen watermark evidence. Equipped with solid theoretical guarantees, **PMark** achieves the desired distortion-free property and improves the robustness against paraphrasing-style attacks. We also provide an empirically optimized version that further removes the requirement for dynamical median estimation for better sampling efficiency. Experimental results show that **PMark** consistently outperforms existing SWM baselines in both text quality and robustness, offering a more effective paradigm for detecting machine-generated text. The source code is available at https://anonymous.4open.science/r/PMark.

Core Task Landscape

This paper addresses: **semantic-level watermarking for large language models**

A total of **50 papers** were analyzed and organized into a taxonomy with **24 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Watermark Generation Mechanisms**
- **Watermark Detection and Verification**
- **Robustness and Attack Resistance**
- **Application-Specific Watermarking**
- **Quality-Aware and Search-Based Optimization**
- **Nested and Hierarchical Watermarking**
- **Survey and Comparative Analysis**

Complete Taxonomy Tree

- semantic-level watermarking for large language models Survey Taxonomy
- Watermark Generation Mechanisms
  - Logit-Based Token Manipulation
  - Green-Red List Partitioning (3 papers)
    - [13] Context-aware watermark with semantic balanced green-red lists for large language models (Yuxuan Guo, 2024) View paper
    - [32] Unbiased Watermark for Large Language Models (Hu, 2023) View paper
    - [33] A Watermark for Large Language Models (Kirchenbauer, 2023) View paper
  - Adaptive Entropy-Aware Watermarking (4 papers)
    - [1] Adaptive text watermark for large language models (Liu, 2024) View paper
    - [3] Scalable watermarking for identifying large language model outputs (Sumanth Dathathri, 2024) View paper
    - [35] Improving the Generation Quality of Watermarked Large Language Models via Word Importance Scoring (Li Yuhang, 2023) View paper
    - [41] Invisible Entropy: Towards Safe and Efficient Low-Entropy LLM Watermarking (Gu Tianle, 2025) View paper
  - Multi-Feature and Dual Watermarking (3 papers)
    - [2] Duwak: Dual watermarks in large language models (Chaoyi Zhu, 2024) View paper
    - [29] Ensemble Watermarks for Large Language Models (Kern, 2024) View paper
    - [30] Bileve: Securing Text Provenance in Large Language Models Against Spoofing with Bi-level Signature (Zhou Tong, 2024) View paper
  - Theoretically Optimized Watermarking (2 papers)
    - [40] Theoretically Grounded Framework for LLM Watermarking: A Distribution-Adaptive Approach (He, 2024) View paper
    - [50] Necessary and Sufficient Watermark for Large Language Models (Takezawa, 2023) View paper
  - Semantic-Level and Sentence-Based Watermarking
  - Semantic Invariance and Proxy Functions ★ (3 papers)
    - [0] PMark: Towards Robust and Distortion-free Semantic-level Watermarking with Channel Constraints (Anon et al., 2026) View paper
    - [4] A robust semantics-based watermark for large language model against paraphrasing (Ren Jie, 2024) View paper
    - [6] A Semantic Invariant Robust Watermark for Large Language Models (Liu Aiwei, 2023) View paper

    - Sentence-Level Similarity and Hashing (3 papers)
      - [5] Simmark: A robust sentence-level similarity-based watermarking algorithm for large language models (Wang Le-le, 2025) View paper
      - [23] CoheMark: A Novel Sentence-Level Watermark for Enhanced Text Quality (Zhang Junyan, 2025) View paper
      - [48] SemBits: Multi-bit Semantic Watermarking with Sentence-Level Hashing for LLMs (Xiangyu Feng, 2025) View paper
    - Topic and Context-Aware Watermarking (2 papers)
      - [14] Topic-Based Watermarks for Large Language Models (Jiang, 2024) View paper
      - [15] Token-specific watermarking with enhanced detectability and semantic coherence for large language models (Huo, 2024) View paper
    - Post-Hoc and Black-Box Watermarking
    - Lexical Substitution and Paraphrasing (3 papers)
      - [9] LOCAT: Localization-driven Text Watermarking via Large Language Models (Liang, 2025) View paper
      - [11] Postmark: A robust blackbox watermark for large language models (Chang, 2024) View paper
      - [38] Robust Multi-bit Text Watermark with LLM-based Paraphrasers (Xu Xiaojun, 2024) View paper
    - Structural and Syntactic Watermarking (2 papers)
      - [8] Post-hoc watermarking for robust detection in text generated by large language models (J Hao, 2025) View paper
      - [25] Signal Watermark on Large Language Models (Xu, 2024) View paper
    - Black-Box Sampling-Based Methods (2 papers)
      - [17] Watermarking text generated by black-box language models (Yang Xi, 2023) View paper
      - [18] Watermarking language models through language models (Agnibh Dasgupta, 2024) View paper
    - Specialized Watermarking Architectures
    - Learning-Based Encoding Modules (3 papers)
      - [16] REMARK-LLM: A Robust and Efficient Watermarking Framework for Generative Large Language Models (Zhang, 2023) View paper
      - [19] DeepTextMark: a deep learning-driven text watermarking approach for identifying large language model generated text (Travis Munyer, 2024) View paper
      - [27] Yet Another Watermark for Large Language Models (BAO Siyuan, 2025) View paper
    - Cross-Attention and Contrastive Mechanisms (2 papers)
      - [21] A Contrastive Semantic Watermarking Framework for Large Language Models (Jianxin Wang, 2025) View paper
      - [22] Cross-Attention watermarking of Large Language Models (Folco Bertini Baldassini, 2024) View paper
    - In-Context and Prompt-Based Watermarking (1 papers)
    - [7] In-Context Watermarks for Large Language Models (Liu, 2025) View paper
- Watermark Detection and Verification
    - Public Verification and Keyless Detection (2 papers)
    - [12] An unforgeable publicly verifiable watermark for large language models (Liu Aiwei, 2023) View paper
    - Likelihood and Statistical Detection (1 papers)
    - [34] A likelihood based approach for watermark detection (X Li, 2025) View paper
    - Semantic Key Modules and Adaptive Detection (2 papers)
    - [24] Robustness Assessment and Enhancement of Text Watermarking for Google's SynthID (Han Xia, 2025) View paper
    - [39] SimKey: A Semantically Aware Key Module for Watermarking Language Models (Kodama, 2025) View paper
- Robustness and Attack Resistance
    - Certified and Provable Robustness (1 papers)
    - [45] A Certified Robust Watermark For Large Language Models (Liu Jian, 2024) View paper
    - Defense Against Spoofing Attacks (1 papers)
    - [26] Defending LLM watermarking against spoofing attacks with contrastive representation learning (An Li, 2025) View paper
- Application-Specific Watermarking
    - Code Generation Watermarking (2 papers)
    - [43] Resilient Watermarking for LLM-Generated Codes (Li Boquan, 2024) View paper
    - [49] Robustness Analysis of Watermarking Techniques for LLM-Generated Code (Di Cao, 2025) View paper
    - Conditional Text Generation Watermarking (1 papers)
    - [31] Watermarking conditional text generation for ai detection: Unveiling challenges and a semantic-aware watermark remedy (Fu Yu, 2024) View paper
    - Model Protection and User Attribution (3 papers)
    - [28] Double-i watermark: Protecting model copyright for llm fine-tuning (Li Shen, 2024) View paper
    - [36] PersonaMark: Personalized LLM watermarking for model protection and user attribution (Zhang Yue-han, 2024) View paper
    - [44] Protecting text ip in the era of llms with robust and scalable watermarking (GKR Lau, 2024) View paper
- Quality-Aware and Search-Based Optimization (1 papers)
    - [42] WaterSearch: A Quality-Aware Search-based Watermarking Framework for Large Language Models (Yukang Lin, 2025) View paper
- Nested and Hierarchical Watermarking (1 papers)
    - [20] A Nested Watermark for Large Language Models (Nagatsuka Koichi, 2025) View paper
- Survey and Comparative Analysis (3 papers)
    - [10] Watermarking for large language models: A survey (Zhiguang Yang, 2025) View paper
    - [46] Watermarking Large Language Models and the Generated Content: Opportunities and Challenges (Ruisi Zhang, 2024) View paper
    - [47] Covert Imprinting: A Comparative Analysis of Watermarking Techniques for Detecting and Attributing Large Language Model Generated Text (ME O'Connel, 2025) View paper

Narrative

Core task: semantic-level watermarking for large language models. The field has organized itself around several complementary branches that address different facets of embedding and verifying invisible signals in LLM-generated text. Watermark Generation Mechanisms explores how to inject marks during or after generation, ranging from token-level biasing to sentence-based semantic transformations that preserve meaning while encoding information. Detection and Verification methods develop statistical tests and learned classifiers to identify watermarked content, while Robustness and Attack Resistance investigates defenses against paraphrasing,

adversarial edits, and spoofing attempts. Application-Specific Watermarking tailors schemes to domains such as code generation or personalized outputs, and Quality-Aware and Search-Based Optimization balances watermark strength against text fluency through search or scoring heuristics. Nested and Hierarchical Watermarking enables multi-level or multi-bit encoding, and Survey and Comparative Analysis provides overviews of the rapidly evolving landscape.

Within the generation mechanisms, a particularly active line of work focuses on semantic-level and sentence-based approaches that modify text at a higher abstraction than individual tokens. PMark[0] exemplifies this direction by introducing proxy functions to maintain semantic invariance during watermark embedding, ensuring that paraphrases or meaning-preserving edits do not erase the signal. This contrasts with earlier token-biasing schemes and aligns closely with Semantic Invariant Watermark[6], which similarly emphasizes preserving semantics through carefully designed transformations. Robust Semantics Watermark[4] also operates in this space, exploring how to achieve both semantic fidelity and resilience to attacks. A central trade-off across these works is between the strength of the watermark signal and the risk of degrading text quality or introducing detectable artifacts, with PMark[0] addressing this challenge through its proxy-based optimization framework that balances imperceptibility and robustness.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. A robust semantics-based watermark for large language model against paraphrasing

Authors: Ren Jie, Jie Ren, Xu Han, Han Xu, Liu Yiding, et al. (16 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Large language models (LLMs) have show great ability in various natural language tasks. However, there are concerns that LLMs are possible to be used improperly or even illegally. To prevent the malicious usage of LLMs, detecting LLM-generated text becomes crucial in the deployment of LLM applications. Watermarking is an effective strategy to detect the LLM-generated content by encoding a pre-defined secret watermark to facilitate the detection process. However, the majority of existing watermar...

Relationship Analysis

Both papers belong to the Semantic Invariance and Proxy Functions category, using semantic embeddings to create watermarks robust to paraphrasing attacks. They overlap in treating sentences as watermarking units and leveraging semantic representations (embeddings) to maintain watermark detectability under text modifications. However, PMark uses cosine similarity with random pivot vectors and multi-channel constraints with median-based sampling for distortion-free generation, while the candidate paper (SemaMark) discretizes high-dimensional embeddings onto a 2D Normalized Embedding Ring and uses contrastive learning to ensure uniform distribution of semantic values for vocabulary partitioning.

---

2. A Semantic Invariant Robust Watermark for Large Language Models

Authors: Liu Aiwei, Pan Leyi, Aiwei Liu, Hu, Xuming, et al. (13 authors total) | **Year/Venue**: 2023 • International Conference on Learning Representations | **URL**: View paper

Abstract

Watermark algorithms for large language models (LLMs) have achieved extremely high accuracy in detecting text generated by LLMs. Such algorithms typically involve adding extra watermark logits to the LLM's logits at each generation step. However, prior algorithms face a trade-off between attack robustness and security robustness. This is because the watermark logits for a token are determined by a certain number of preceding tokens; a small number leads to low security robustness, while a large ...

Relationship Analysis

Both papers belong to the Semantic Invariance and Proxy Functions category, using semantic embeddings to create watermarks robust to paraphrasing attacks. They overlap in their core approach of leveraging semantic-level features (embeddings) rather than token-level patterns to achieve robustness against text modifications. However, the original paper (PMark) focuses on distortion-free watermarking through dynamic median estimation and multi-channel constraints with theoretical guarantees, while the candidate paper (SIR) trains a specialized watermark model to transform semantic embeddings into watermark logits, emphasizing the trade-off between attack robustness and security robustness.

Contributions Analysis

Overall novelty summary. The paper proposes PMark, a semantic-level watermarking method built on a theoretical framework of proxy functions that map sentences to scalar values. It sits within the 'Semantic Invariance and Proxy Functions' leaf of the taxonomy, which contains only three papers total. This represents a relatively sparse research direction compared to more crowded areas like token-level logit manipulation, suggesting the paper targets a less saturated niche focused on sentence-based watermarking with formal semantic guarantees.

The taxonomy reveals that semantic-level watermarking divides into three main approaches: proxy functions (this paper's leaf), sentence-level similarity/hashing, and topic-aware methods. Neighboring leaves address related but distinct challenges—similarity-based methods leverage embeddings for detection, while topic-aware techniques incorporate contextual information. The paper's proxy function framework appears to bridge theoretical optimization (a separate branch with formal guarantees) and semantic robustness, positioning it at the intersection of multiple research threads within the generation mechanisms category.

Among eighteen candidates examined across three contributions, the multi-channel constraint mechanism shows the most substantial prior overlap, with two refutable candidates identified from eight examined. The theoretical proxy function framework and the PMark method itself appear more novel, with zero refutable candidates among nine and one examined respectively. This suggests the core formalism and implementation may be relatively fresh, while the idea of using multiple constraints for robustness has closer precedents in the limited search scope.

Based on the top-eighteen semantic matches examined, the work appears to introduce a distinct theoretical angle within a sparsely populated research direction. The analysis does not cover the full breadth of watermarking literature, and the small candidate pool means potentially relevant work outside the semantic search radius may exist. The taxonomy structure indicates this is an emerging area with room for novel contributions, though the multi-channel mechanism overlaps with existing robustness strategies.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Theoretical framework for semantic-level watermarking via proxy functions

Description: The authors introduce a theoretical framework that unifies existing semantic-level watermarking methods through the concept of proxy functions—functions that map sentences to scalar values. This framework provides analytical foundations for evaluating watermarking performance and enables formal analysis of distortion and robustness properties.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Deep model intellectual property protection via deep watermarking

URL: View paper

Brief Assessment

Deep Model Protection[52] focuses on protecting deep neural networks for image processing tasks through spatial invisible watermarking, not on semantic-level text watermarking. The candidate paper does not address proxy functions for sentence-level watermarking in LLMs.

---

2. Noisy But Forgotten: LLM Unlearning are Robust against Perturbed Data in the Wild

URL: View paper

Brief Assessment

Noisy But Forgotten[56] focuses on LLM unlearning robustness against perturbed data, not on watermarking theory. The candidate paper examines how data perturbations (incomplete, rewritten, watermarked) affect unlearning performance, which is a completely different research problem from developing theoretical frameworks for semantic-level watermarking.

---

3. Watermarking language models through language models

URL: View paper

Brief Assessment

Watermarking through LMs[18] focuses on prompt-driven watermarking using instruction-following capabilities of LLMs, not on semantic-level watermarking with proxy functions that map sentences to scalar values for distribution analysis.

---

4. Universally optimal watermarking schemes for llms: from theory to practice

URL: View paper

Brief Assessment

Universally Optimal Schemes[54] focuses on token-level watermarking with a hypothesis testing framework for independence between text and auxiliary variables, not semantic-level watermarking using proxy functions that map sentences to scalar values.

---

5. Theoretically Grounded Framework for LLM Watermarking: A Distribution-Adaptive Approach

URL: View paper

Brief Assessment

Distribution-Adaptive Framework[40] focuses on token-level watermarking with distribution-adaptive approaches for LLMs, not semantic-level watermarking using proxy functions as fundamental units for sentences.

---

6. DRSW: Dual-stage Robust Semantic Watermarking for Image Semantic Communication

URL: View paper

Brief Assessment

DRSW[51] focuses on image semantic communication watermarking using transform-domain and deep learning methods. The candidate does not address text-based semantic watermarking or proxy function frameworks for LLMs.

---

7. On Forging Semantic Watermarks in Diffusion Models: A Theoretical Perspective

URL: View paper

Brief Assessment

Forging Semantic Watermarks[57] focuses on semantic watermarks in diffusion models for image generation, analyzing forgery attacks through rate-distortion theory. The ORIGINAL paper addresses semantic-level watermarking for large language models (LLMs) using proxy functions that map sentences to scalar values. These are fundamentally different domains (image vs. text generation) with distinct technical approaches and objectives.

---

8. Model watermarking for image processing networks

URL: View paper

Brief Assessment

Model Watermarking Networks[53] focuses on protecting image processing models through spatial invisible watermarking in model outputs, not on theoretical frameworks for semantic-level text watermarking using proxy functions that map sentences to scalar values.

---

9. Token-Specific Watermarking with Enhanced Detectability and Semantic Coherence for Large Language Models. 2024 International Conference on Machine â¦

URL: View paper

Brief Assessment

Token-Specific Enhanced[55] focuses on token-level watermarking with semantic coherence optimization, not semantic-level watermarking frameworks using proxy functions for sentence-level operations.

Contribution 2: Multi-channel constraint mechanism for enhanced robustness

Description: The authors identify that sparse watermark evidence in existing semantic-level watermarking methods weakens robustness against attacks. They address this by introducing multiple channel constraints (using orthogonal pivot vectors) to increase the density of watermark evidence, thereby improving robustness against paraphrasing and word-level attacks.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Improved unbiased watermark for large language models

URL: View paper

Prior Art Analysis

Improved Unbiased[62] demonstrates that multi-channel mechanisms for improving robustness in watermarking were proposed prior to the original paper. The candidate introduces MCMark, which partitions vocabulary into multiple segments (channels) and promotes token probabilities within selected segments to enhance detectability and robustness. This directly addresses the same problem of sparse watermark evidence that the original paper claims to solve through multi-channel constraints. Both papers use orthogonal or independent channels to increase watermark evidence density, and both demonstrate improved robustness against paraphrasing and modification attacks through this multi-channel approach.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly address the problem of improving robustness through multiple channels/segments and provide theoretical analysis of the detectability-robustness trade-off. - **Original**: we identify that sparse watermarking evidence in swms negatively impacts adversarial robustness, and address this problem by introducing multiple channel constraints. - **Candidate**: we theoretically demonstrate that when the number of segments equals two, mcmark offers superior detectability compared to dipmark(wu et al., 2023) and sta-1 (mao et al., 2024). we further discuss the trade-offs between detectability and robustness in mcmark.

Evidence 2 - **Rationale**: Both papers analyze how multiple channels affect robustness against attacks. The candidate provides explicit mathematical analysis of robustness trade-offs with channel numbers, demonstrating prior work on this concept. - **Original**: multi-channel sampling. given a pre-defined set of b orthogonal vectors v1, ··· , vb, we refer to each pivot as achannel. the proxy function defined onvi is fvi(s) = ⟨vi, t (s)⟩, also denoted asfi for simplicity. - **Candidate**: an adversary may attempt to alter the output token to disrupt the watermark detection. in mcmark detection, if a token xt is modified to x′ t and x′ t /∈ vit, the watermark signal is effectively removed. consequently, the probability that a watermark is removed due to such an alteration is given by ...

---

2. Robust Watermarks Leak: Channel-Aware Feature Extraction Enables Adversarial Watermark Manipulation

URL: View paper

Brief Assessment

Robust Watermarks Leak[60] focuses on attacking watermarking systems by exploiting multi-channel feature extraction to remove or forge watermarks, not on designing multi-channel constraints to improve watermark robustness. The paper's multi-channel approach serves an adversarial purpose rather than a watermark design contribution.

---

3. Proactive Deepfake Detection via Self-Verifiable Semantic Watermarking

URL: View paper

Brief Assessment

Proactive Deepfake Detection[63] focuses on embedding watermarks in facial images for deepfake detection, not on semantic watermarking for text generation. The multi-channel approach in the original paper applies to text-based LLM watermarking with orthogonal pivot vectors, while the candidate addresses image-based facial semantic features.

---

4. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification

URL: View paper

Brief Assessment

Ringid[61] focuses on multi-channel watermarking for diffusion models in image generation, not semantic-level text watermarking. The technical domains and applications are fundamentally different.

---

5. Peccavi: Visual Paraphrase Attack Safe and Distortion Free Image Watermarking Technique for AI-Generated Images

URL: View paper

Brief Assessment

Peccavi[67] focuses on multi-channel frequency domain watermarking for images to resist visual paraphrase attacks, not on semantic-level text watermarking with channel constraints to improve robustness against paraphrasing attacks in LLM outputs.

---

6. Pattern-based quantum text watermarking: Securing digital content with next-Gen quantum techniques

URL: View paper

Brief Assessment

Pattern-based Quantum[64] focuses on quantum-based image watermarking techniques using multi-channel quantum images (mcqi), not semantic-level text watermarking with orthogonal pivot vectors for LLM-generated content.

---

7. Multi-Channel Statistical Framework for Robust and Reliable Watermark Detection in Color Image Processing

URL: View paper

Brief Assessment

Multi-Channel Statistical[66] focuses on RGB channel dependencies in color image watermarking using HMMs, not semantic-level text watermarking with orthogonal pivot vectors for LLM-generated content.

---

8. Words are not enough: sentence level natural language watermarking

URL: View paper

Prior Art Analysis

Words Not Enough[65] demonstrates prior work on using multiple channels to enhance watermarking robustness in natural language. The candidate explicitly describes using word-based methods as a separate channel from sentence-based methods to improve resilience and embedding properties. This multi-channel approach predates the original paper's contribution, showing that the concept of using multiple channels (orthogonal features) to increase watermark evidence density was already established in semantic watermarking literature.

Evidence

Evidence 1 - **Rationale**: The candidate paper explicitly mentions exploiting orthogonality between features in their sentence-level watermarking technique, which is the same fundamental concept (orthogonal constraints) that the original paper uses for their multi-channel mechanism. - **Original**: we introduce multiple channel constraints to enhance the density of watermarking evidence, as illustrated in figure 1. leveraging the fact that random vectors in high-dimensional spaces are almost always orthogonal - **Candidate**: the sentence level watermarking technique we introduce is novel and powerful, as it relies on multiple features of each sentence and exploits the notion of orthogonality between features.

Contribution 3: PMark: distortion-free semantic watermarking with online and offline variants

Description: The authors propose PMark, a semantic-level watermarking method with two variants: an online version that dynamically estimates the proxy function median and is theoretically distortion-free, and an offline version that uses a prior median assumption (zero) to reduce computational cost while maintaining low distortion. Both variants enforce multiple channel constraints to strengthen watermark evidence.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Zero watermarking for text on www using semantic approach

URL: View paper

Brief Assessment

Zero Watermarking WWW[59] focuses on semantic watermarking for web text using a different technical approach. The candidate does not demonstrate that distortion-free semantic watermarking with online/offline computational variants existed prior to PMark.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] PMark: Towards Robust and Distortion-free Semantic-level Watermarking with Channel Constraints View paper
- [1] Adaptive text watermark for large language models View paper
- [2] Duwak: Dual watermarks in large language models View paper
- [3] Scalable watermarking for identifying large language model outputs View paper
- [4] A robust semantics-based watermark for large language model against paraphrasing View paper
- [5] Simmark: A robust sentence-level similarity-based watermarking algorithm for large language models View paper
- [6] A Semantic Invariant Robust Watermark for Large Language Models View paper
- [7] In-Context Watermarks for Large Language Models View paper
- [8] Post-hoc watermarking for robust detection in text generated by large language models View paper
- [9] LOCAT: Localization-driven Text Watermarking via Large Language Models View paper
- [10] Watermarking for large language models: A survey View paper
- [11] Postmark: A robust blackbox watermark for large language models View paper
- [12] An unforgeable publicly verifiable watermark for large language models View paper
- [13] Context-aware watermark with semantic balanced green-red lists for large language models View paper
- [14] Topic-Based Watermarks for Large Language Models View paper
- [15] Token-specific watermarking with enhanced detectability and semantic coherence for large language models View paper
- [16] REMARK-LLM: A Robust and Efficient Watermarking Framework for Generative Large Language Models View paper
- [17] Watermarking text generated by black-box language models View paper
- [18] Watermarking language models through language models View paper
- [19] DeepTextMark: a deep learning-driven text watermarking approach for identifying large language model generated text View paper
- [20] A Nested Watermark for Large Language Models View paper
- [21] A Contrastive Semantic Watermarking Framework for Large Language Models View paper
- [22] Cross-Attention watermarking of Large Language Models View paper
- [23] CoheMark: A Novel Sentence-Level Watermark for Enhanced Text Quality View paper
- [24] Robustness Assessment and Enhancement of Text Watermarking for Google's SynthID View paper
- [25] Signal Watermark on Large Language Models View paper
- [26] Defending LLM watermarking against spoofing attacks with contrastive representation learning View paper
- [27] Yet Another Watermark for Large Language Models View paper
- [28] Double-i watermark: Protecting model copyright for llm fine-tuning View paper
- [29] Ensemble Watermarks for Large Language Models View paper
- [30] Bileve: Securing Text Provenance in Large Language Models Against Spoofing with Bi-level Signature View paper
- [31] Watermarking conditional text generation for ai detection: Unveiling challenges and a semantic-aware watermark remedy View paper
- [32] Unbiased Watermark for Large Language Models View paper
- [33] A Watermark for Large Language Models View paper
- [34] A likelihood based approach for watermark detection View paper
- [35] Improving the Generation Quality of Watermarked Large Language Models via Word Importance Scoring View paper
- [36] PersonaMark: Personalized LLM watermarking for model protection and user attribution View paper
- [37] A Private Watermark for Large Language Models View paper
- [38] Robust Multi-bit Text Watermark with LLM-based Paraphrasers View paper
- [39] SimKey: A Semantically Aware Key Module for Watermarking Language Models View paper
- [40] Theoretically Grounded Framework for LLM Watermarking: A Distribution-Adaptive Approach View paper
- [41] Invisible Entropy: Towards Safe and Efficient Low-Entropy LLM Watermarking View paper
- [42] WaterSearch: A Quality-Aware Search-based Watermarking Framework for Large Language Models View paper
- [43] Resilient Watermarking for LLM-Generated Codes View paper
- [44] Protecting text ip in the era of llms with robust and scalable watermarking View paper
- [45] A Certified Robust Watermark For Large Language Models View paper
- [46] Watermarking Large Language Models and the Generated Content: Opportunities and Challenges View paper
- [47] Covert Imprinting: A Comparative Analysis of Watermarking Techniques for Detecting and Attributing Large Language Model Generated Text View paper
- [48] SemBits: Multi-bit Semantic Watermarking with Sentence-Level Hashing for LLMs View paper
- [49] Robustness Analysis of Watermarking Techniques for LLM-Generated Code View paper
- [50] Necessary and Sufficient Watermark for Large Language Models View paper
- [51] DRSW: Dual-stage Robust Semantic Watermarking for Image Semantic Communication View paper
- [52] Deep model intellectual property protection via deep watermarking View paper
- [53] Model watermarking for image processing networks View paper
- [54] Universally optimal watermarking schemes for llms: from theory to practice View paper
- [55] Token-Specific Watermarking with Enhanced Detectability and Semantic Coherence for Large Language Models. 2024 International Conference on Machine â¦ View paper
- [56] Noisy But Forgotten: LLM Unlearning are Robust against Perturbed Data in the Wild View paper
- [57] On Forging Semantic Watermarks in Diffusion Models: A Theoretical Perspective View paper
- [58] Latents-Inv: Robust Semantic Watermark with Key-Assisted Recovery for diffusion models View paper
- [59] Zero watermarking for text on www using semantic approach View paper

- [60] Robust Watermarks Leak: Channel-Aware Feature Extraction Enables Adversarial Watermark Manipulation View paper
- [61] Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification View paper
- [62] Improved unbiased watermark for large language models View paper
- [63] Proactive Deepfake Detection via Self-Verifiable Semantic Watermarking View paper
- [64] Pattern-based quantum text watermarking: Securing digital content with next-Gen quantum techniques View paper
- [65] Words are not enough: sentence level natural language watermarking View paper
- [66] Multi-Channel Statistical Framework for Robust and Reliable Watermark Detection in Color Image Processing View paper
- [67] Peccavi: Visual Paraphrase Attack Safe and Distortion Free Image Watermarking Technique for AI-Generated Images View paper