Novelty Assessment Report

Paper: Break the Trade-off Between Watermark Strength and Speculative Sampling Efficiency for Language Models

PDF URL: https://openreview.net/pdf?id=HA8vzzT6Ax

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Watermarking is a principled approach for tracing the provenance of large language model (LLM) outputs, but its deployment in practice is hindered by inference inefficiency. Speculative sampling accelerates inference, with efficiency improving as the acceptance rate between draft and target models increases. Yet recent work reveals a fundamental trade-off: higher watermark strength reduces acceptance, preventing their simultaneous achievement. We revisit this trade-off and show it is not absolute. We introduce a quantitative measure of watermark strength that governs statistical detectability and is maximized when tokens are deterministic functions of pseudorandom numbers. Using this measure, we fully characterize the trade-off as a constrained optimization problem and derive explicit Pareto curves for two existing watermarking schemes. Finally, we introduce a principled mechanism that injects pseudorandomness into draft-token acceptance, ensuring maximal watermark strength while maintaining speculative sampling efficiency. Experiments further show that this approach improves detectability without sacrificing efficiency. Our findings uncover a principle that unites speculative sampling and watermarking, paving the way for their efficient and practical deployment.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Watermarking Language Models with Speculative Sampling Acceleration**

A total of **6 papers** were analyzed and organized into a taxonomy with **6 categories**.

Taxonomy Overview

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

- **Watermarking-Acceleration Trade-off Analysis**
- **Watermarking Implementation Methods**
- **Watermark Security and Robustness**
- **Theoretical Foundations of Machine Learning**

Complete Taxonomy Tree

- Watermarking Language Models with Speculative Sampling Acceleration Survey Taxonomy
- Watermarking-Acceleration Trade-off Analysis
  - Theoretical Trade-off Characterization (1 papers)
  - [2] Inevitable Trade-off between Watermark Strength and Speculative Sampling Efficiency for Language Models (Hu, 2024) View paper
  - Trade-off Resolution Methods ★ (2 papers)
  - [0] Break the Trade-off Between Watermark Strength and Speculative Sampling Efficiency for Language Models (Anon et al., 2026) View paper
  - [3] Watermarking using Semantic-aware Speculative Sampling: from Theory to Practice (B Huang, n.d.) View paper
- Watermarking Implementation Methods
  - Production-Scale Watermarking Systems (1 papers)
  - [1] Scalable watermarking for identifying large language model outputs (Sumanth Dathathri, 2024) View paper
  - Advanced Watermarking Capabilities (1 papers)
  - [6] SAEMark: Steering Personalized Multilingual LLM Watermarks with Sparse Autoencoders (Z Yu, n.d.) View paper
- Watermark Security and Robustness (1 papers)
  - [5] An Experimental Study on Attacks and Vulnerabilities of Text Watermarks (Albrecht, 2025) View paper
- Theoretical Foundations of Machine Learning (1 papers)
  - [4] Understanding and Enhancing Machine Learning Models with Theoretical Foundations (Hu, 2024) View paper

Narrative

Core task: watermarking language models with speculative sampling acceleration. The field addresses the challenge of embedding detectable signals into LLM outputs while maintaining generation speed through speculative decoding techniques. The taxonomy organizes work into several main branches: Watermarking-Acceleration Trade-off Analysis examines the inherent tension between watermark strength and inference efficiency; Watermarking Implementation Methods covers practical embedding schemes and detection algorithms; Watermark Security and Robustness investigates resilience against adversarial attacks and text modifications; and Theoretical Foundations of Machine Learning provides the mathematical underpinnings. Representative works like Scalable LLM Watermarking[1] and Inevitable Watermark Tradeoff[2] establish fundamental constraints, while studies such as Text Watermark Attacks[5] probe security boundaries. The branches interconnect around the central question of whether watermarking and acceleration can coexist without compromising either objective.

A particularly active line explores trade-off resolution methods, seeking to reconcile watermark detectability with speculative sampling speedups that traditionally interfere with embedding schemes. Watermark Speculative Tradeoff[0] sits squarely within this branch, addressing how speculative decoding's draft-verify mechanism can disrupt watermark consistency. It shares thematic concerns with Semantic Speculative Watermarking[3], which similarly navigates the interplay between acceleration and signal preservation, though the two may differ in their specific technical approaches or semantic constraints. Meanwhile, works like SAEMark[6] explore alternative embedding strategies that might sidestep certain acceleration conflicts. The original paper's emphasis on resolving this trade-off positions it among efforts to make watermarking practical for production systems where both provenance tracking and low-latency generation are essential, contrasting with purely theoretical analyses or security-focused studies that treat acceleration as secondary.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Watermarking using Semantic-aware Speculative Sampling: from Theory to Practice

Authors: B Huang, H Zhu, J Piet, B Zhu, JD Lee | **URL**: View paper

Abstract

â¦ understanding of watermarking in large language models (â¦ -aware Speculative Sampling), a novel watermarking algorithm â¦ use of maximal coupling via speculative sampling allows it to â¦

Relationship Analysis

Both papers belong to the Trade-off Resolution Methods category, addressing the fundamental tension between watermark strength and speculative sampling efficiency in language models. They share overlapping goals of achieving strong watermarking while maintaining high sampling efficiency, both proposing mechanisms that leverage pseudorandomness in the acceptance process. The key difference is that the original paper focuses on breaking the trade-off through pseudorandom draft-token acceptance with theoretical characterization via Pareto curves, while the candidate paper (SEAL) approaches the problem through semantic-aware random seeds and maximal coupling via speculative sampling, with emphasis on statistical limits and practical tamper resistance.

Contributions Analysis

Overall novelty summary. The paper addresses the watermark-acceleration trade-off in language models by proposing a pseudorandom draft-token acceptance mechanism. It resides in the 'Trade-off Resolution Methods' leaf, which contains only two papers including this one. This sparse population suggests the specific problem of reconciling watermark strength with speculative sampling efficiency remains relatively underexplored. The taxonomy shows six total papers across six leaf nodes, indicating the broader field of watermarking with acceleration is still emerging rather than saturated.

The taxonomy places this work within 'Watermarking-Acceleration Trade-off Analysis', adjacent to 'Theoretical Trade-off Characterization' and separate from 'Watermarking Implementation Methods'. The sibling paper in the same leaf likely explores similar resolution strategies, while neighboring leaves address theoretical constraints or production-scale deployment without acceleration concerns. The scope notes clarify that this branch focuses on breaking or optimizing trade-offs, distinguishing it from pure theoretical analysis or security evaluations found elsewhere in the taxonomy structure.

Among fifteen candidates examined, the quantitative watermark strength measure shows one refutable candidate out of four examined, suggesting some prior conceptualization exists. The constrained optimization characterization examined ten candidates with none refuting, indicating potential novelty in formalizing the trade-off mathematically. The pseudorandom acceptance mechanism examined only one candidate with no refutation, though the limited search scope means undiscovered prior work could exist. The statistics reflect a focused semantic search rather than exhaustive coverage, leaving room for undetected overlaps.

Based on the limited search of fifteen candidates, the work appears to occupy a relatively sparse research direction with modest prior overlap. The single refutable contribution among three analyzed suggests incremental advancement on watermark strength formalization, while the optimization framework and acceptance mechanism show no clear precedent within the examined scope. However, the small candidate pool and emerging field structure mean a broader literature review could reveal additional related efforts.

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

Contribution 1: Quantitative measure of watermark strength

Description: The authors propose a continuous measure of watermark strength based on expected KL divergence, which quantifies how strongly tokens depend on pseudorandomness. This measure governs the decay rate of p-values in detection and is maximized when tokens are deterministic functions of pseudorandom numbers.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. SemBits: Multi-bit Semantic Watermarking with Sentence-Level Hashing for LLMs

URL: View paper

Brief Assessment

SemBits[7] appears to use KL divergence in a different context (comparing text generated directly from the language model). The limited context provided does not demonstrate that SemBits[7] proposed the same quantitative measure of watermark strength based on expected KL divergence governing p-value decay rates.

2. Fast segmentation of watermarked texts from large language models through epidemic change-points framework

URL: View paper

Brief Assessment

Fast Watermark Segmentation[8] focuses on localizing watermarked segments within texts using epidemic change-point detection, not on quantifying watermark strength via KL divergence or analyzing its relationship to detection efficiency.

3. Towards Better Statistical Understanding of Watermarking LLMs

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

4. Learnable Linguistic Watermarks for Tracing Model Extraction Attacks on Large Language Models

URL: View paper

Prior Art Analysis

Learnable Linguistic Watermarks[9] demonstrates that prior work exists using KL divergence as a quantitative measure of watermark strength. The candidate paper explicitly defines watermark strength in terms of KL divergence and uses it to govern statistical detectability, showing that this approach predates the original paper's contribution. Both papers use KL divergence to quantify how strongly tokens depend on pseudorandomness and link this measure to detection efficiency through hypothesis testing.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly define watermark strength using KL divergence and establish that this measure governs detection efficiency. - **Original**: we introduce a quantitative measure of watermark strength for unbiased watermarks, defined as the expected kl divergence between the watermarked and original token distributions. we show that this measure governs the decay rate of p-values - **Candidate**: that also proves that the kl divergence stands for the strength of the watermark. the large kl divergence means a stronger watermark intensity

Evidence 2 - **Rationale**: Both papers use expected KL divergence to quantify the difference between watermarked and original distributions, establishing this as the core measure of watermark strength. - **Original**: Definition 3.1. for a watermarking scheme that samples tokens from the modified distribution pζ = s(p, ζ), its watermark strength is defined as ws(pζ) = eζ[dkl(pζ ∥p)] = eζ [x w∈w pζ,w log pζ,w pw] - **Candidate**: i ˆplm - ipa = ewi~ ˆplm [-log(ˆplm(wi) pa(wi))] = kl(ˆplm(wi)||pa(wi)) > 0

Evidence 3 - **Rationale**: Both papers connect KL divergence-based watermark strength to hypothesis testing and p-value decay, demonstrating the same theoretical framework for linking strength to detectability. - **Original**: theorem 3.1 (sample complexity via p-value decay). let α ∈ (0, 1) and w1:n = (w1, . . . , wn). consider the hypothesis testing problem based on n independent samples - **Candidate**: considering the type i error rate not greater than α, we can define the upper bound bi to accept hypothesis h0 as : p r([w1, w2, ...] ~ pa; i ˆplm (h0) - ipa(h1) < bi) < α

Evidence 4 - **Rationale**: Both papers derive explicit bounds relating the number of samples needed for detection to the KL divergence measure, showing the same mathematical relationship between watermark strength and sample complexity. - **Original**: in particular, to guarantee thep-value ≤ α, it is necessary that n ≥ 1 d log 1 α (1 + o(1)) - **Candidate**: thus the information difference bound bi and the upper bound of type i error rate α should satisfy the following equation: bi ≥ -ln(α 1 - α)

Contribution 2: Characterization of the trade-off as constrained optimization

Description: The authors formalize the trade-off between watermark strength and sampling efficiency as a Pareto frontier problem. They provide explicit optimization formulations and derive trade-off curves for existing watermarking methods including Gumbel-max and SynthID.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Dual secure robust watermarking scheme based on hybrid optimization algorithm for image security

URL: View paper

Brief Assessment

Dual Secure Watermarking[12] focuses on image watermarking using hybrid optimization for robustness and security, not on language model watermarking or the trade-off between watermark strength and sampling efficiency in LLMs.

2. Adaptor: Improving the robustness and imperceptibility of watermarking by the adaptive strength factor

URL: View paper

Brief Assessment

Adaptor[13] addresses trade-offs in image watermarking (strength factor vs. image quality/robustness), not the specific trade-off between watermark strength and sampling efficiency in language models that the original paper formalizes as constrained optimization.

3. Optimal Watermark Generation under Type I and Type II Errors

URL: View paper

Brief Assessment

Optimal Watermark Generation[15] formulates watermarking as a hypothesis testing problem with type I/II error constraints, focusing on f-divergence minimization. The original paper addresses a different problem: the trade-off between watermark strength and speculative sampling efficiency in LLMs, formulated as a Pareto frontier optimization.

4. Optimized Dynamic Watermarking for Audio DNNs with Adaptive Embedding and Boundary Sampling

URL: View paper

Brief Assessment

Audio DNN Watermarking[17] focuses on audio signal watermarking for DNN model protection, not on language model watermarking or speculative sampling efficiency trade-offs.

5. Inevitable Trade-off between Watermark Strength and Speculative Sampling Efficiency for Language Models

URL: View paper

Brief Assessment

Inevitable Watermark Tradeoff[2] focuses on proving impossibility results (no-go theorem) for simultaneous optimization, while the original paper formulates the trade-off as a solvable constrained optimization problem with explicit Pareto curves. The approaches are complementary rather than overlapping.

6. Stereo robust watermark algorithm based on parameter optimization

URL: View paper

Brief Assessment

Stereo Robust Watermark[16] addresses audio watermarking with parameter optimization for imperceptibility-robustness trade-offs, not the trade-off between watermark strength and sampling efficiency in language models formalized as constrained optimization.

7. Optimization of Multibit Watermarking

URL: View paper

Brief Assessment

Multibit Watermark Optimization[18] focuses on optimizing multibit watermarking for robustness and transparency in spread spectrum methods, not on the trade-off between watermark strength and speculative sampling efficiency in language models.

8. Improving the performance of DCT-based fragile watermarking using intelligent optimization algorithms

URL: View paper

Brief Assessment

DCT Fragile Watermarking[19] focuses on optimizing DCT-based fragile watermarking parameters using genetic algorithms for image authentication, not on the trade-off between watermark strength and sampling efficiency in language models.

9. Adversarially Robust Digital Watermarking via Data-Centric Optimization

URL: View paper

Brief Assessment

Adversarially Robust Watermarking[14] focuses on adversarial robustness of digital watermarks in images via data-centric optimization, not on the trade-off between watermark strength and sampling efficiency in language models as a constrained optimization problem.

10. Robin: Robust and invisible watermarks for diffusion models with adversarial optimization

URL: View paper

Brief Assessment

Robin[11] addresses watermarking for diffusion models (image generation), not language models. The paper focuses on balancing watermark robustness and invisibility in generated images through adversarial optimization, which is a fundamentally different domain and problem formulation than the ORIGINAL paper's focus on watermark strength versus speculative sampling efficiency in language model inference.

Contribution 3: Pseudorandom draft-token acceptance mechanism

Description: The authors propose a novel mechanism that makes the acceptance decision in speculative sampling pseudorandom rather than truly random. This approach achieves maximal watermark strength while preserving sampling efficiency, breaking the previously established trade-off.

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. An Experimental Study on Attacks and Vulnerabilities of Text Watermarks

URL: View paper

Brief Assessment

Text Watermark Attacks[5] focuses on evaluating attacks against existing watermarking systems (KGW, X-SIR, SynthID) rather than proposing novel mechanisms for speculative sampling or watermark embedding. The candidate does not address pseudorandom draft-token acceptance in speculative sampling.

Appendix: Text Similarity Detection

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

References

- [0] Break the Trade-off Between Watermark Strength and Speculative Sampling Efficiency for Language Models View paper
- [1] Scalable watermarking for identifying large language model outputs View paper
- [2] Inevitable Trade-off between Watermark Strength and Speculative Sampling Efficiency for Language Models View paper
- [3] Watermarking using Semantic-aware Speculative Sampling: from Theory to Practice View paper
- [4] Understanding and Enhancing Machine Learning Models with Theoretical Foundations View paper
- [5] An Experimental Study on Attacks and Vulnerabilities of Text Watermarks View paper
- [6] SAEMark: Steering Personalized Multilingual LLM Watermarks with Sparse Autoencoders View paper
- [7] SemBits: Multi-bit Semantic Watermarking with Sentence-Level Hashing for LLMs View paper
- [8] Fast segmentation of watermarked texts from large language models through epidemic change-points framework View paper
- [9] Learnable Linguistic Watermarks for Tracing Model Extraction Attacks on Large Language Models View paper
- [10] Towards Better Statistical Understanding of Watermarking LLMs View paper
- [11] Robin: Robust and invisible watermarks for diffusion models with adversarial optimization View paper
- [12] Dual secure robust watermarking scheme based on hybrid optimization algorithm for image security View paper
- [13] Adaptor: Improving the robustness and imperceptibility of watermarking by the adaptive strength factor View paper
- [14] Adversarially Robust Digital Watermarking via Data-Centric Optimization View paper
- [15] Optimal Watermark Generation under Type I and Type II Errors View paper
- [16] Stereo robust watermark algorithm based on parameter optimization View paper
- [17] Optimized Dynamic Watermarking for Audio DNNs with Adaptive Embedding and Boundary Sampling View paper
- [18] Optimization of Multibit Watermarking View paper
- [19] Improving the performance of DCT-based fragile watermarking using intelligent optimization algorithms View paper