# Poster: SemAnchor: Enhancing Extraction Accuracy of Multi-bit LLM Text Watermarking via Semantic-Guided Segment Assignment

Zhiwei Xu[†*], Xiaolin Peng[†*], Jiabao Gao[†], Dinghong Song[‡], Tian Qiu[†], Hai Wan[†] and Xibin Zhao[†§]
[†]*KLISS, BNRist, School of Software, Tsinghua University, China*
[‡]*Department of Electrical Engineering and Computer Science, University of California, Merced*

***Abstract*—Unlike zero-bit watermarking, which can only provide binary detection, multi-bit text watermarking has attracted growing interest due to its ability to attribute watermark bits to outputs of Large Language Models (LLMs). Currently, multi-bit watermarking typically embeds watermark bits into the partitioned token space with hash-based assignment mechanisms. However, this approach would introduce a critical limitation termed *Synchronization Vulnerability* – semantically identical tokens assigned to divergent segments can easily lead to watermark extraction failures.**

**In this work, we propose SEMANCHOR, a novel multi-bit LLM text watermarking framework that incorporates semantic cluster anchoring into bit assignment. By exploiting the LLM's intrinsic embedding space through a two-stage semantic balanced partitioning strategy, Semantic Atomization and Frequency-Aware Packing, SEMANCHOR locks semantically interchangeable tokens to the same watermark segment. Evaluations on LLaMA-2-7B demonstrate that SEMANCHOR achieves >97% accuracy while exhibiting superior resilience to semantic perturbations compared to state-of-the-art methods.**

## 1. Introduction

The proliferation of LLMs necessitates robust content provenance mechanisms to mitigate misuse like disinformation [2], [4]. While zero-bit watermarking [2] solely offers binary detection, multi-bit schemes provide granular traceability [6]. To overcome the computationally expensive enumeration of earlier schemes [6], recent state-of-the-art (SOTA) approaches have shifted towards hash-based assignment [4], [7] to balance accuracy and time complexity.

Unfortunately, we identify a critical *Synchronization Vulnerability* in the prevailing hash-based assignment approach: semantically identical tokens (e.g., "happy" vs. "glad") mapped to divergent segments could break the alignment between tokens and watermark segments, easily causing decoding failures, as illustrated in Fig. 1.

To address this vulnerability, we propose SEMANCHOR, a novel multi-bit LLM text watermarking framework that incorporates robust semantic-level assignment. SEMANCHOR employs a two-stage semantic balanced partitioning strategy, including Semantic Atomization to build anchors and Frequency-Aware Packing to balance load. Experiments on OpenGen [3] benchmark with LLaMA-2-7B [5] demonstrate the effectiveness and robustness of SEMANCHOR, achieving $> 97\%$ accuracy with inherent stability against semantic attacks. Our contributions are twofold:

- **New Vulnerability Identification:** We identify a vulnerability that significantly weakens the performance of existing multi-bit approaches.
- **Semantically Robust Framework:** We propose SEMANCHOR, addressing the vulnerability by anchoring watermark segments to semantic clusters.

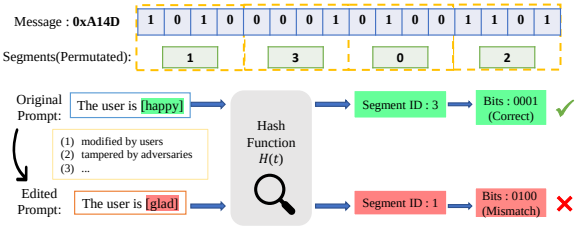



Figure 1: The Synchronization Vulnerability. A single synonym substitution ("happy" → "glad") can disrupt segment alignment, leading to decoding failures.

## 2. SEMANCHOR Design

We propose two-stage semantic balanced partitioning to *reconcile* semantic consistency for robustness with segment-level uniformity for balance. This strategy resolves the Synchronization Vulnerability while overcoming the token frequency imbalance identified in prior work [4].

### Stage 1: Semantic Atomization for Robustness

Instead of brittle hash mappings [4], [7], we project the vocabulary into LLM's native embedding space. We partition tokens into fine-grained "Semantic Atoms" (micro-clusters), ensuring synonyms are locked to the same Cluster ID. This creates anchors for robust bit assignment.

*Equal contribution.
§Corresponding author (zxb@tsinghua.edu.cn).

TABLE 1: Performance comparison on Match Rate (the proportion of generated text that can exactly extract embedded watermark message without error) and Bit Accuracy (the ratio of correctly extracted message bits without error-correction codes) across different bit lengths.

| Method | 16 Bits | | 20 Bits | | 24 Bits | | 32 Bits | |
|---|---|---|---|---|---|---|---|---|
| | Match Rate | Bit Accuracy | Match Rate | Bit Accuracy | Match Rate | Bit Accuracy | Match Rate | Bit Accuracy |
| Yoo et al. [7] | 73.6 | 94.6 | 49.2 | 90.8 | 30.4 | 89.4 | 8.4 | 81.2 |
| Cohen et al. [1] | 88.8 | 97.0 | 78.4 | 95.3 | 65.6 | 94.7 | 27.2 | 89.7 |
| Qu et al. [4] | 98.0 | 99.7 | 97.6 | 99.6 | 96.0 | 99.5 | 94.0 | 99.1 |
| **SEMANCHOR (ours)** | **99.0** | **99.8** | **99.0** | **99.8** | **99.0** | **99.6** | **97.0** | **99.2** |

### Stage 2: Frequency-Aware Packing for Balance

Pure semantic clustering leads to a highly unbalanced token distribution, causing sparse segment coverage and extraction failures. To mitigate this, we formulate assignment as a *Bin Packing Problem*. We employ a Frequency-Aware Greedy Packing algorithm to achieve near-uniform probability mass, ensuring that every watermark segment has an equal probability of being activated.[1]

## 3. Evaluation

**Experimental Setup.** We evaluated SEMANCHOR on the OpenGen [3] benchmark using LLaMA-2-7B [5], and conducted a comparative analysis with SOTAs [1], [4], [7].

**How effective is SEMANCHOR?** As shown in Table 1, SEMANCHOR consistently outperforms all baselines across all evaluated bit lengths. Specifically, it maintains $> 97.0\%$ Match Rate for 16-32 bit lengths and $> 99.2\%$ Bit Accuracy overall, demonstrating significant improvements over existing multi-bit methods.

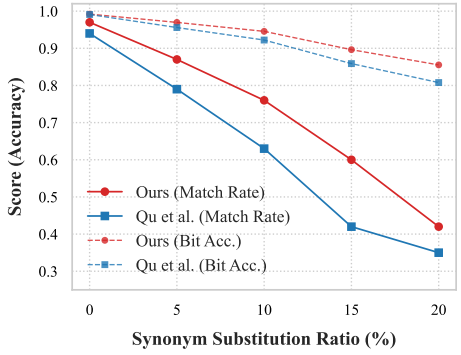



Figure 2: Robustness under Synonym Substitution Attacks.

**Why is SEMANCHOR effective?** Fig. 2 shows resilience against synonym substitution attacks. SEMANCHOR exhibits a more promising degradation curve than the SOTA [4], achieving an 87% Match Rate at 5% perturbation (vs. 79%) and preserving a 7% accuracy lead under severe 20% distortion. This confirms that SEMANCHOR successfully mitigates the Synchronization Vulnerability where SOTAs fail.

1. Stages 1 & 2 are pre-computed offline. Online text generation only requires an $O(1)$ table lookup with negligible cost.

## 4. Conclusion

We identify the Synchronization Vulnerability in existing multi-bit watermarking approaches. By reconciling semantic consistency for robustness with segment-level uniformity for balance, SEMANCHOR successfully addresses the Synchronization Vulnerability with a novel semantic-guided segment assignment mechanism. Evaluations on LLaMA-2-7B demonstrate that SEMANCHOR achieves state-of-the-art accuracy (>97%) and superior robustness against semantic-level attacks. Future work will focus on defending against structural multi-word paraphrasing and scaling to modern LLMs with expanded vocabularies.

## Acknowledgments

This work is in part supported by the Guangdong S&T Programme (No. 2024B0101030002). We sincerely thank all referees for their valuable efforts for this work.

## References

[1] Aloni Cohen, Alexander Hoover, and Gabe Schoenbach. Watermarking language models for many adaptive users. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 2583–2601. IEEE, 2025.

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

[3] Kalpesh Krishna, Yixiao Song, Marzena Karpinska, et al. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 2023.

[4] Wenjie Qu, Wengrui Zheng, Tianyang Tao, Dong Yin, Yanze Jiang, et al. Provably robust multi-bit watermarking for ai-generated text. In *34th USENIX Security Symposium (USENIX Security 25)*.

[5] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[6] Lean Wang, Wenkai Yang, Deli Chen, Hao Zhou, Yankai Lin, Fandong Meng, Jie Zhou, and Xu Sun. Towards codable watermarking for injecting multi-bits information to llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024*.

[7] KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. Advancing beyond identification: Multi-bit watermark for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

# SemAnchor: Enhancing Extraction Accuracy of Multi-bit LLM Text Watermarking via Semantic-Guided Segment Assignment

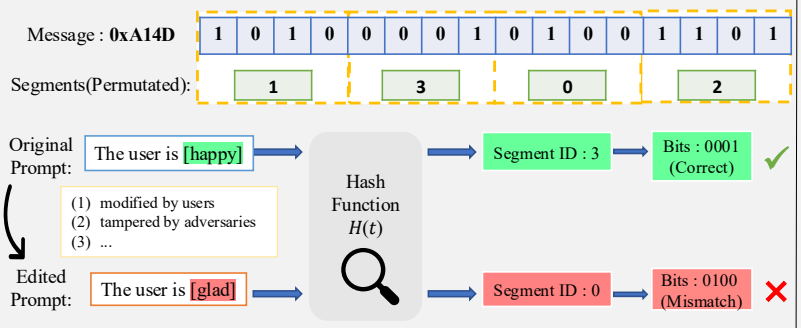
Zhiwei Xu[1*], Xiaolin Peng[1*], Jiabao Gao[1], Dinghong Song[2], Tian Qiu[1], Hai Wan[1], Xibin Zhao[1(✉)]
[1]Tsinghua University [2]University of California, Merced [*]Equal contribution
Contact: {zxb@tsinghua.edu.cn}

## Introduction

- **Background**: Reliable content tracing calls for effective and robust **Multi-bit LLM text watermarking**.
- **Challenges:** Current SOTA schemes rely on Hash-based Assignment, which could suffer from a "**Synchronization Vulnerability**".

## The "Synchronization Vulnerability"

- **Problem:** Synonym substitutions disrupt watermark segment alignment.




- **Problem:** Synonym substitutions disrupt watermark segment alignment.
- **Effect:** Slightly edited text causes catastrophic mismatch.

## Our Insight

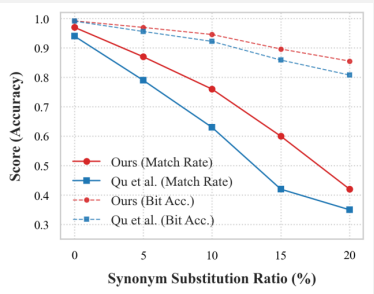
**"Reconciling Semantic Consistency (robustness) with Segment-Level Uniformity (balance)"**

- **Principle:** Semantically interchangeable tokens should anchor to the same watermark segment to maintain semantic consistency.

## Evaluations

**Setup**: LLaMA-2-7B on OpenGen Benchmark.

| Method | 16 Bits | | 20 Bits | | 24 Bits | | 32 Bits | |
|---|---|---|---|---|---|---|---|---|
| | Match Rate | Bit Accuracy | Match Rate | Bit Accuracy | Match Rate | Bit Accuracy | Match Rate | Bit Accuracy |
| Yoo et al. [11] | 73.6 | 94.6 | 49.2 | 90.8 | 30.4 | 89.4 | 8.4 | 81.2 |
| Cohen et al. [1] | 88.8 | 97.0 | 78.4 | 95.3 | 65.6 | 94.7 | 27.2 | 89.7 |
| Qu et al. [8] | 98.0 | 99.7 | 97.6 | 99.6 | 96.0 | 99.5 | 94.0 | 99.1 |
| SEMANCHOR (ours) | **99.0** | **99.8** | **99.0** | **99.8** | **99.0** | **99.6** | **97.0** | **99.2** |

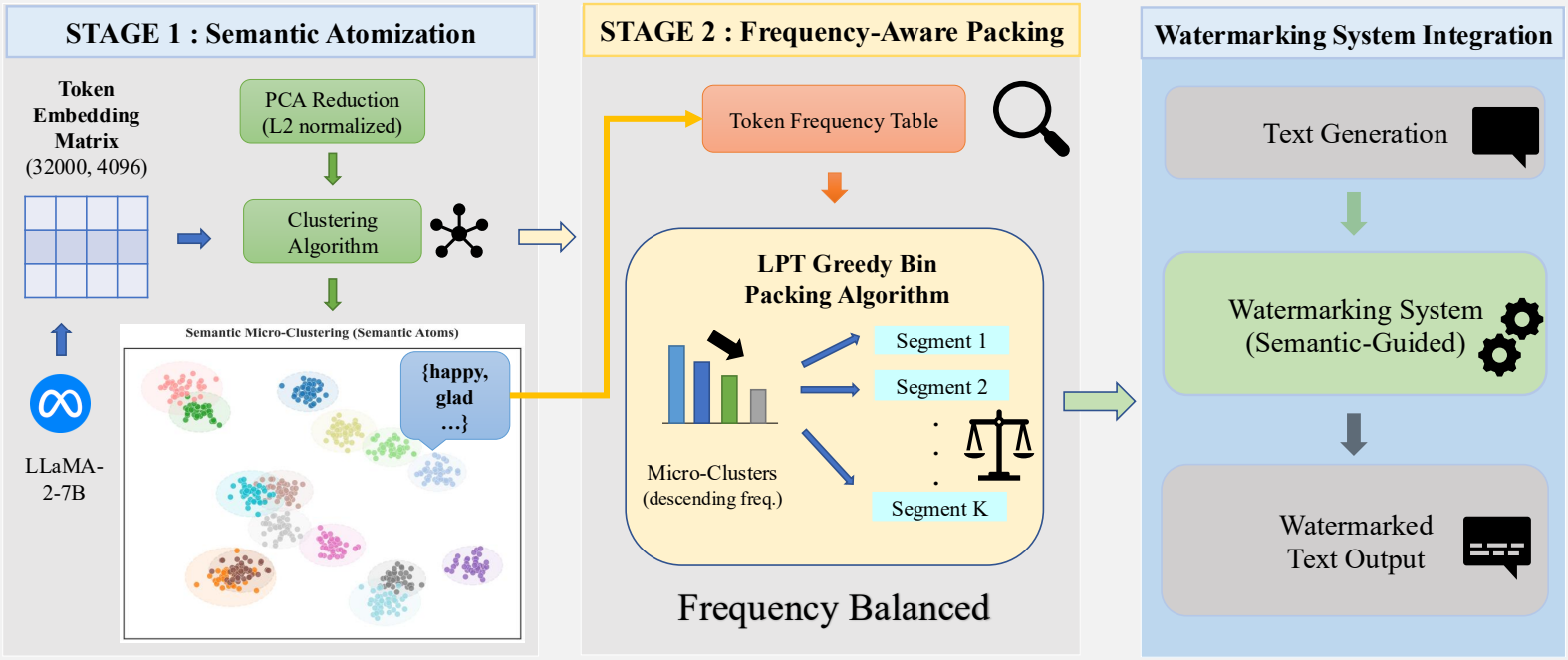



**Key Results**:

- Consistently achieves SOTA accuracy (> 97%) across all payloads.
- Maintains an 87% Match Rate at 5% perturbation against synonym substitution attacks.

## Our Solution: SemAnchor Framework

We propose a **Two-Stage Semantic Balanced Partitioning** strategy:




- **Stage 1: Semantic Atomization**
  **Process:** Partition tokens into Micro-Clusters in the embedding space.
  **Effect:** $\boldsymbol{Token\ t \approx Token\ t' \Rightarrow Cluster(t) = Cluster(t')}$
- **Stage 2: Frequency-Aware Packing**
  **Challenge:** Zipfian long-tail Distribution leads to imbalance.
  **Solution:** Bin Packing Problem with Greedy LPT.