International Journal of Engineering Research & Science (IJOER) ISSN: [2395-6992] [Vol-12, Issue-3, March 2026]

A Cross-Model Evaluation of the KGW Watermarking Method across LLMs: A Systematic Review

Rushabh Patil[1*]; Harsh Mali[2]; Ankit Purohit[3]; Kirtida Naik[4]

Department of Computer Engineering, University of Mumbai, India
*Corresponding Author

Abstract— *The rapid growth of open-source large language models (LLMs) has increased the need for reliable methods to verify the origin of machine-generated text. Statistical watermarking has emerged as a promising solution by embedding imperceptible signals into generated outputs while preserving semantic quality. Among existing approaches, the Green-List/Red-List (KGW) watermarking method has gained significant attention due to its practicality and strong detectability under controlled conditions. This paper presents a comprehensive review and comparative analysis of KGW watermarking across recent research studies involving open-source LLMs. Existing methodologies, robustness evaluations, attack strategies, and alternative watermarking techniques are systematically examined and summarized. The comparative analysis highlights variations in watermark effectiveness across model architectures and identifies key limitations related to robustness under adversarial transformations. The study provides insights into current research trends, unresolved challenges, and future directions for developing reliable watermarking systems in open-source LLM environments.*

Keywords— Large Language Models, Text Watermarking, KGW Watermark, AI Content Provenance, Open-Source Models, Robustness Analysis, Watermark Attacks, AI Security.

I. INTRODUCTION

Large Language Models (LLMs) have demonstrated remarkable capabilities in generating coherent, context-aware, and human-like text, leading to their widespread adoption across applications such as automated content creation, virtual assistants, and educational platforms. However, the increasing accessibility of LLMs has raised serious concerns related to misinformation, plagiarism, copyright misuse, and the inability to reliably distinguish AI-generated content from human-authored text. Ensuring trustworthy provenance mechanisms for machine-generated text has therefore become a critical research requirement.

Text watermarking has emerged as a promising solution by embedding imperceptible statistical signals into generated text that can later be detected to verify origin and authenticity. The first formal decoding-time watermarking approach for LLMs was introduced by Kirchenbauer et al. [1], demonstrating that biased token selection can enable reliable statistical detection while maintaining text fluency. Subsequent research has extended this concept to explore robustness in open-source environments, where users have direct access to model parameters and inference pipelines [2].

Recent studies have shown that watermark signals can be significantly weakened through paraphrasing, fine-tuning, translation, and black-box scrubbing attacks, raising concerns about real-world deployment reliability [4], [8], [10], [12]. These findings indicate that while watermarking remains a promising provenance tool, its robustness must be carefully evaluated under realistic threat models.

This paper presents a comprehensive review and comparative analysis of the KGW watermarking method across recent research on open-source LLMs, highlighting robustness trends, limitations, attack strategies, and emerging alternatives.

II. LITERATURE SURVEY

Kirchenbauer et al. [1] — This foundational paper introduced the first decoding-based watermarking algorithm for LLMs, establishing a framework for statistical signal embedding in generated text. The approach leverages the Hugging Face Transformers Logits Processor to adjust token probabilities during generation, enforcing a green-list/red-list partitioning of the vocabulary. Tokens from the green list are biased by an additive logit shift (δ) while the overall ratio of biased tokens is controlled by a sampling fraction (γ). Through experiments on OPT-family models, the authors showed that with parameters $\delta \in [0.5, 2.0]$ and $\gamma \approx 0.25$, the watermark remains imperceptible to human readers while being highly detectable using statistical hypothesis testing.

Gloaguen et al. [2] — This study extended the evaluation of watermarking durability in the context of open-source models, where users can directly access and modify weights. The authors conducted a systematic analysis of the KGW watermark (δ=2, γ=0.25) under white-box modifications such as fine-tuning on new datasets, model quantization, and parameter merging across checkpoints. Their experiments revealed that while watermark signatures are initially detectable, these common practices significantly degrade or erase the watermark signal, reducing statistical confidence in detection.

Chen et al. [3] — Building on the challenges identified in prior work, this paper introduced WAPITI, a train-free watermarking strategy that embeds watermarks at the parameter level rather than exclusively during decoding. Unlike KGW, which perturbs token probabilities dynamically, WAPITI integrates a watermark vector directly into model parameters through low-rank adaptation techniques. This allows the watermark to survive downstream fine-tuning, as it becomes an intrinsic part of the model's learned representation space.

Huang et al. [4] — This paper proposed B⁴, a universal watermark scrubbing attack that operates under a black-box threat model. Unlike white-box removal methods, B⁴ assumes no prior knowledge of the watermarking algorithm, its structure, or its hyperparameters. The attack is formulated as a constrained optimization problem, where the objective is to reduce the statistical detectability of the watermark while maintaining semantic fidelity in the generated text.

Aremu et al. [5] — This research investigated a novel attack vector against watermarking systems, termed "watermark stealing." The study showed that proprietary watermark rules in closed-source APIs can be reverse-engineered through systematic querying of a watermarked model. Once stolen, these rules can be used to either forge fake watermarks or strip existing ones, undermining both authenticity verification and copyright protection.

Liu et al. [6] — This paper introduced a semantic-invariant watermarking method that aims to resist paraphrasing and semantic-preserving edits. Instead of relying solely on random green/red list partitions, the approach leverages a dedicated embedding LLM to generate semantic vector representations of the context. A trained watermarking model then uses these representations to guide token-level perturbations, determining the green/red split in a manner that aligns with semantic invariance.

Chen et al. [7] — This research presented MCMARK, a family of unbiased watermarks designed to improve both quality and detectability in models such as LLAMA-3. Unlike traditional biased watermarking methods that skew the output distribution, MCMARK partitions the vocabulary into multiple disjoint segments (l), each associated with a unique watermarking channel. During generation, the model selects tokens across channels without introducing significant bias into the probability distribution, thereby preserving text fluency and diversity.

He et al. [8] — This study investigated the resilience of text watermarks under cross-lingual translation, a common obfuscation tactic. The authors introduced the Cross-lingual Watermark Removal Attack (CWRA), where watermarked text is translated into a pivot language and then translated back into the original language. Experimental results showed that current watermarking methods, including KGW, failed to maintain their statistical signatures after round-trip translation.

Pan et al. [9] — This paper introduced MARKLLM, the first comprehensive open-source framework for watermarking research. The toolkit provides standardized implementations of watermarking algorithms, visualization tools for statistical analysis, and configurable evaluation metrics for robustness and quality. By offering an extensible architecture, MARKLLM enables researchers to plug in new watermarking and detection algorithms for direct comparison.

Chang et al. [10] — This research proposed the Smoothing Attack, a watermark removal method based on selectively modifying low-confidence tokens. The intuition is that watermark-carrying tokens often coincide with positions where the model is less confident, and replacing them with higher-confidence alternatives reduces the watermark signal. The method uses a weaker reference model to identify candidate positions and applies smoothing operations that preserve fluency.

An et al. [11] — This paper addressed spoofing attacks, where malicious edits retain the watermark but distort semantics. The authors proposed a defense based on contrastive representation learning, which trains the watermark to be invariant to semantic-preserving edits while being sensitive to semantic-distorting modifications. Experiments demonstrated that this method improved robustness to spoofing while maintaining detection reliability.

Cheng et al. [12] — This work introduced the Self-Information Rewrite Attack (SIRA), which exploits token-level entropy to target watermark removal. The intuition is that watermarking algorithms often embed their strongest signals in high-

entropy tokens, which carry greater uncertainty. By selectively rewriting these tokens, SIRA effectively erases watermark signatures while maintaining fluency and semantic coherence.

Li et al. [13] — This research explored watermarking at the parameter level using weight quantization. The method embeds watermark signals directly into quantized weight values, making them intrinsic to the model rather than applied during generation. Because the watermark is embedded at the weight level, it becomes more durable against common adversarial attacks such as fine-tuning or paraphrasing.

Xu et al. [14] — This paper proposed a sentence-level watermarking framework that embeds multi-bit watermarks using LLM-based paraphrasers as encoders and a text classifier as the decoder. Training was performed with Proximal Policy Optimization (PPO)-based reinforcement learning to balance fidelity and detectability. The approach achieved extremely high detection accuracy (AUC > 99.99%) and demonstrated robustness against common perturbations such as synonym replacement and light paraphrasing.

Huo et al. [15] — This work introduced a multi-objective optimization approach that assigns token-specific watermark parameters. Instead of applying a fixed γ (green-list ratio) and δ (logit bias) across all tokens, the system uses lightweight neural networks to generate dynamic values for each token, guided by SimCSE embeddings. This ensures that the watermark is both detectable and semantically coherent, avoiding unnatural token biases.

III. ANALYSIS TABLE

TABLE 1
SUMMARY OF WATERMARKING METHODS AND ATTACKS

Sr. No.	Paper Title (Year)	Technology Used	Key Hyperparameters	Main Contribution/Result
1	A Watermark for Large Language Models (2023)	OPT model, Hugging Face Transformers, LogitsProcessor	δ (0.5–2.0), γ (≈0.25)	First decoding-time watermarking framework balancing imperceptibility with detectability
2	Towards Watermarking of Open-Source LLMs (2025)	Open-Source Models, KGW watermark distillation	δ=2, γ=0.25, k=1	Analysis of durability against fine-tuning, quantization, and model merging
3	WAPITI: A Watermark for Finetuned Open-Source LLMs (2024)	Parameter integration, watermark vector	KGW: δ∈{1,2}, γ=0.25; AAR: k∈{2,3,4}	Train-free defense against fine-tuning attacks
4	B⁴: A Black-Box Scrubbing Attack on LLM Watermarks (2025)	Black-box optimization, constrained optimization	Not specified	Universal scrubbing attack requiring no prior knowledge of watermarking algorithm
5	Watermark Stealing (2025)	API querying, reverse-engineering	Cost <$50 per million tokens	Reverse-engineering watermark rules via API queries
6	Semantic Invariant Robust Watermark (2023)	Embedding LLM, trained watermark model	k1=20, k2=1000, λ1=10, λ2=0.1	Attack robustness and security robustness simultaneously
7	Improved Unbiased Watermark (2025)	MCMARK, LLAMA-3	l (number of segments)	Mitigates text quality trade-off while maintaining output distribution
8	Can Watermarks Survive Translation? (2024)	CWRA, pivot languages	Not specified	Cross-lingual consistency analysis of watermarking technologies
9	MARKLLM: Open-Source Toolkit (2024)	Unified framework, evaluation tools	Configurable parameters	Standardized research enabling systematic comparison
10	Watermark Smoothing Attacks (2024)	Low-confidence token replacement	δ, τ, γ	Exploits relationship between model confidence and watermark detectability
11	Defending Against Spoofing Attacks (2025)	Contrastive representation learning	δ=0.13, Entropy Threshold=2.0, Temperature=0.7	Watermark sensitive to semantic-distorting changes but insensitive to paraphrasing

12	Self-Information Rewrite Attacks (2025)	SIRA	Not specified	Targets high-entropy tokens for high success rates at low cost
13	Watermarking with Weight Quantization (2024)	Weight quantization	Not specified	Embedding watermarks directly into model weights for IP protection
14	Robust Multi-bit Text Watermark (2022)	LLM-based paraphrasers, PPO-RL	λw, λs, λk	Multi-bit watermark at sentence level with >99.99% detection accuracy
15	Token-Specific Watermarking (2024)	MOO, SimCSE embeddings, lightweight networks	Dynamic γ and δ	Dynamically balances detectability and semantic integrity per token

IV. RESEARCH GAP

Although existing research demonstrates that KGW watermarking provides strong detectability under controlled decoding conditions, most studies either focus on a single model family or evaluate robustness primarily under adversarial transformations such as fine-tuning, paraphrasing, and black-box attacks. Limited attention has been given to systematically comparing KGW performance across multiple modern open-source LLM architectures under identical generation and detection settings. Furthermore, variations in watermark signal strength and statistical separation across model families remain underexplored. This creates a gap in understanding how model architecture influences watermark retention and real-world reliability.

Note on Title-Paper Alignment: The title of this paper references a "cross-model evaluation," suggesting original experimental comparison of KGW across different LLM architectures. However, the current manuscript is structured as a literature review. To address this gap, the authors should either: (a) conduct an original experimental study comparing KGW performance across models (e.g., Llama, Mistral, OPT, Gemma) under identical conditions, or (b) revise the title to reflect that this is a systematic review. The present revision adopts the latter approach with the subtitle "A Systematic Review."

V. CRITICAL ANALYSIS

The existing body of research on text watermarking for large language models demonstrates that statistical watermarking techniques can effectively embed detectable provenance signals under controlled generation conditions [1]. However, multiple studies indicate that decoding-time watermarking methods such as KGW suffer from limited robustness when exposed to common post-generation transformations, including paraphrasing, translation, fine-tuning, and knowledge distillation [2], [8], [10], [12].

Furthermore, advanced black-box attacks have shown that watermark detectability can be significantly reduced without requiring knowledge of the watermarking algorithm or model parameters, posing a serious threat to real-world deployment [4], [10], [12]. These attacks exploit token-level uncertainty, confidence smoothing, and semantic rewriting strategies to suppress statistical signals while preserving text fluency.

In response to these vulnerabilities, alternative approaches such as parameter-level watermarking and semantic-invariant watermarking have been proposed to enhance durability against adversarial manipulation [3], [6], [13]. While these methods demonstrate improved robustness, they introduce increased computational complexity, additional training requirements, and potential scalability limitations. Moreover, many existing studies evaluate watermarking techniques on limited model families or restricted datasets, raising concerns regarding generalization across diverse architectures and real-world applications.

Overall, the literature suggests that current watermarking approaches remain insufficient as standalone solutions for robust provenance protection in open-source LLM ecosystems.

VI. FUTURE RESEARCH DIRECTIONS

Future research should focus on developing **hybrid watermarking frameworks** that combine decoding-time, semantic-aware, and parameter-level approaches to improve robustness across diverse adversarial threat models [3], [13]. **Adaptive**

watermarking strategies that dynamically adjust embedding strength based on contextual entropy and token uncertainty may further enhance resistance against black-box scrubbing attacks [10], [12].

Additionally, broader empirical evaluation across multilingual, domain-specific, and long-form generation scenarios is necessary to assess real-world deployment readiness [8]. Establishing **standardized benchmarks**, open-source evaluation toolkits, and reproducible protocols will be critical for advancing watermarking research toward secure and practical large-scale adoption [9].

Specific recommendations for future work:

- Cross-model evaluation of KGW under identical experimental conditions
- Development of robustness metrics for watermarking systems
- Integration of watermarking with model watermarking for multi-layer provenance

VII. CONCLUSION

This review highlights the evolving landscape of text watermarking techniques developed to address provenance, misuse, and intellectual property concerns in large language models. Early decoding-time approaches, particularly the Green-List/Red-List (KGW) watermarking scheme, demonstrated that statistical signals can be embedded into generated text while maintaining fluency and semantic quality. Subsequent studies extended this foundation by examining watermark robustness in open-source environments, revealing that common practices such as fine-tuning, knowledge distillation, paraphrasing, and translation can significantly weaken watermark detectability, although residual signals may still persist under certain conditions.

Comparative analysis across the surveyed literature indicates that watermark effectiveness varies considerably depending on model architecture, decoding behavior, and parameter selection. While KGW remains a practical baseline solution, recent research has shown that advanced black-box scrubbing attacks and watermark stealing techniques pose substantial challenges to decoding-time watermarking. In response, alternative strategies such as parameter-level watermarking, semantic-invariant methods, and token-adaptive approaches have been proposed to improve resilience against adversarial manipulation.

Overall, the literature suggests that **no single watermarking method currently provides comprehensive protection across all threat models**. Decoding-based schemes are effective for initial provenance verification but are insufficient as standalone defenses in adversarial or open-source settings. Future research should prioritize hybrid watermarking strategies, standardized evaluation frameworks, and improved robustness against black-box attacks to support secure and sustainable deployment of large language models.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to their supervisor and guide for invaluable guidance, continuous support, and encouragement throughout this research. The authors also acknowledge their institution for providing the necessary facilities and a conducive research environment. They further appreciate the constructive feedback and suggestions from faculty members and colleagues, which significantly enhanced the quality of this work. The authors are deeply grateful to their families and friends for their unwavering support and motivation.

CONFLICT OF INTEREST

The authors declare no conflict of interest regarding the publication of this paper

REFERENCES

[1] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)* (pp. 17061–17084).

[2] Gloaguen, T., Jovanović, N., Staab, R., & Vechev, M. (2025). Towards watermarking of open-source LLMs. In *1st Workshop on GenAI Watermarking, collocated with ICLR 2025*. https://arxiv.org/abs/2502.10525

[3] Chen, Z., Li, J., Chen, R., & Zeng, Z. (2024). WAPITI: A watermark for finetuned open-source LLMs. *arXiv*, arXiv:2410.06467.

[4] Huang, B., Pu, X., & Wan, X. (2025). B⁴: A black-box scrubbing attack on LLM watermarks. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)* (pp. 9113–9126).

[5] Jovanović, N., Staab, R., & Vechev, M. (2024). Watermark stealing in large language models. In *International Conference on Machine Learning (ICML) 2024*. https://arxiv.org/abs/2402.19361

[6] Liu, A., Pan, L., Hu, X., Meng, S., & Wen, L. (2023). A semantic invariant robust watermark for large language models. In *International Conference on Learning Representations (ICLR) 2024*. *https://arxiv.org/abs/2310.06356*

[7] Chen, R., Wu, Y., Guo, J., & Huang, H. (2025). Improved unbiased watermark for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 20587–20601).

[8] He, Z., Zhou, B., Hao, H., Liu, A., Wang, X., Tu, Z., Zhang, Z., & Wang, R. (2024). Can watermarks survive translation? On the cross-lingual consistency of text watermark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 4115–4129).

[9] Pan, L., Hu, X., Li, X., Wu, Y., Liu, S., He, Z., & Wen, L. (2024). MARKLLM: An open-source toolkit for LLM watermarking. In *EMNLP 2024 Demo. https://arxiv.org/abs/2405.10051*

[10] Chang, H., Liu, S., & Miers, I. (2024). Watermark smoothing attacks against language models. *arXiv*, arXiv:2407.14206.

[11] An, L., Liu, Y., Liu, Y., Zhang, Y., Bu, Y., & Chang, S. (2025). Defending LLM watermarking against spoofing attacks with contrastive representation learning. *arXiv*, arXiv:2504.06575.

[12] Cheng, Y., Guo, H., Li, Y., & Sigal, L. (2025). Revealing weaknesses in text watermarking through self information rewrite attacks. In *International Conference on Machine Learning (ICML) 2025. https://arxiv.org/abs/2505.05190*

[13] Li, X., Grandvalet, Y., & Davoine, F. (2023). Watermarking LLMs with weight quantization. *arXiv*, arXiv:2310.11237.

[14] Xu, G., Zhang, D., & Chen, J. (2024). Robust multi-bit text watermark with LLM-based paraphrasers. In *ICML 2025. https://arxiv.org/abs/2412.03123*

[15] Huo, M., Zhang, S., Yang, J., & Wang, Q. (2024). Token-specific watermarking with enhanced detectability and semantic coherence for large language models. *arXiv*, arXiv:2402.18059.