**RESEARCH PAPER**

# Entropy-Gated Sparse Watermarking for Robust Provenance Detection of Paraphrased LLM-Generated Text

Aryan Uniyal[1*], Vivek Kumar Tamta[2] and Papendra Kumar[3]

*[1,2,3]Department of Computer Science and Engineering, Govind Ballabh Pant Institute of Engineering & Technology, Pauri Garhwal, Uttarakhand 246194, India.*
*Email Id: [1]aryanuniyal25@gmail.com (A.U.), [2]vivek.tamta2010@gmail.com (V.K.T.) and [3]papendra1@gmail.com (P.K.)*
*[*]**Corresponding Author:** Aryan Uniyal*
*Orcid id: https://orcid.org/0009-0006-1053-6521.*

***Received: 28th Feb, 2026; Revised: 6th March 2026; Accepted: 7th April, 2026; Available Online: 20th April, 2026***

**ABSTRACT**
Reliable provenance for large language model (LLM)-generated text becomes difficult once generated content is paraphrased, translated, or otherwise rewritten before audit. We evaluate adaptive entropy-gated sparse watermarking (AEG) as a generation-time signal for information provenance under controlled false-positive calibration. The configuration uses a reduced greenlist fraction, $\gamma = 0.25$, and concentrates watermark bias at high-entropy token positions, where generation choices are less predetermined. Detection thresholds are estimated empirically from clean model outputs, enabling true-positive rate (TPR) comparison at a 1% false-positive rate (FPR) operating point against a Kirchenbauer–Geiping–Wen (KGW) baseline. Across LLaMA-3-8B, Mistral-7B, Qwen2-7B, and BLOOM-7B1 with N = 300 samples per tested condition, AEG improves post-paraphrase detection over KGW in three of four model regimes. The strongest result appears for Qwen2-7B, where oracle-minimum TPR at 1% FPR rises from 17.3% to 98.8%; LLaMA-3-8B and Mistral-7B show smaller supporting gains, while BLOOM-7B1 favors KGW. The reversal in BLOOM-7B1 argues against a simple tokenizer-size explanation and points instead to model-specific vocabulary-regime effects. Together, these results support calibrated, model-aware watermark evaluation for LLM-generated text provenance, especially when content may be rewritten across distributed AI and digital publishing workflows before audit.

**Keywords**: *large language model watermarking; AI-generated text provenance; information security; watermark robustness; paraphrase robustness; entropy-gated watermarking; empirical calibration; false-positive control; generative AI.*
**How to cite this article:** Uniyal A, Tamta VK, Kumar P. Entropy-Gated Sparse Watermarking for Robust Provenance Detection of Paraphrased LLM-Generated Text. Int J Drug Deliv Technol. 2026;16(63s):1844-1858. DOI: 10.25258/ijddt.16.63s.192
**Source of support:** Nil.
**Conflict of interest:** None

## 1. INTRODUCTION

Large language model (LLM)-generated text now moves through information workflows in which source, authenticity, and accountability matter. The same passage may be generated by a controlled model, revised by a user, paraphrased, translated, or republished before it is audited. In this setting, provenance is not just a raw classification problem. A useful detector must retain evidence after ordinary or adversarial rewriting while keeping the false-positive rate low enough for downstream use in document drafting, moderation, automated reporting, and digital publishing workflows.

Generation-time watermarking offers one way to support this kind of audit. A keyed signal is inserted while the model generates text, and a later detector tests the final token sequence for that signal. In the greenlist watermarking family introduced by Kirchenbauer et al., a secret key partitions the vocabulary into green and red tokens, generation is biased toward green tokens, and detection uses a z-score based on the observed green-token count. The practical question is whether such a signal remains measurable after paraphrase and backtranslation, especially when the same watermark configuration is transferred across models.

This paper studies that question as a model-aware information-provenance problem. The experiments evaluate watermark robustness under controlled rewriting attacks rather than a full deployment of distributed AI pipelines, so the results should be read as evidence for a provenance component rather than an end-to-end system. A fixed greenlist fraction can behave differently across LLMs because the nominal tokenizer size is not the same as the set of tokens a model actively uses for English generation. The effective operating regime is shaped by

**Author for Correspondence: aryanuniyal25@gmail.com*

the tokenizer, the training mixture, and language coverage. We therefore ask whether a sparser greenlist, combined with entropy-aware injection, can concentrate the watermark where token choice is less constrained and improve detection after rewriting.

We evaluate a sparse entropy-gated configuration within the Kirchenbauer–Geiping–Wen (KGW)-style generation-time watermarking family. The greenlist fraction is reduced from $\gamma = 0.50$ to $\gamma = 0.25$, and an Adaptive Entropy Gate (AEG) applies watermark bias mainly at high-entropy positions. The evaluation is empirical and comparative: it tests whether this combined sparse-gated configuration improves post-rewrite detection, while also identifying model regimes in which it does not dominate the baseline.

The study uses four open-weight LLMs at N = 300 samples per condition: LLaMA-3-8B (128K tokens), Mistral-7B (32K tokens), Qwen2-7B (152K tokens), and BLOOM-7B1 (251K tokens). The largest gain appears on Qwen2-7B, where oracle-minimum true-positive rate at 1% false-positive rate rises from 17.3% for KGW to 98.8% for AEG. LLaMA-3-8B and Mistral-7B show supporting gains, whereas BLOOM-7B1 favors KGW. This negative case is important: it argues against treating nominal tokenizer size as a sufficient explanation for post-paraphrase watermark robustness.

The main contributions are:

- We evaluate entropy-gated sparse watermarking for provenance detection of paraphrased or rewritten LLM-generated text at a controlled empirical false-positive operating point.
- We provide cross-model evidence that post-paraphrase robustness depends on how dense the greenlist is within each model's actively used English vocabulary, rather than on nominal tokenizer size alone.
- We report paired tests, ablations, and margin diagnostics showing that the observed gains are tied to the combined sparse-gated configuration and that deployment still requires model-specific threshold calibration, key screening, and stronger adaptive-attack evaluation.

## 2. RELATED WORK

Kirchenbauer et al. (2023) introduced a practical LLM watermarking scheme in which a keyed procedure partitions the vocabulary into greenlist and redlist tokens, biases generation toward green tokens, and detects the resulting signal with a z-score test. Kirchenbauer et al. (2024) later showed that a stronger logit bias can improve reliability, although it may also affect text quality. The present study remains within this greenlist-based detection family. Its contribution is not a new detector statistic; it asks how a sparse, entropy-gated injection policy behaves after the watermarked text is rewritten.

Several related watermarking approaches make different tradeoffs. Christ et al. (2023) study distribution-preserving schemes designed to avoid detectable distortion, whereas Kuditipudi et al. (2023) focus on robustness to insertions and deletions. These methods address important parts of the design space, but they differ from the provenance-audit setting considered here, where a lightweight z-score detector is applied to rewritten text without access to generation logits at audit time. Multi-bit schemes, including those proposed by Yoo et al. (2023) and Hou et al. (2023), encode payloads rather than a binary provenance signal, creating a separate capacity–robustness tradeoff. This paper keeps the binary green/red partition and measures how much of that signal survives paraphrase attacks.

Rewrite attacks are a central stress test for watermarking. Krishna et al. (2023) showed that DIPPER-style paraphrasing can evade many AI-text detectors, including watermark-based detectors. Zhao et al. (2023) proposed SIR, a context-dependent watermarking baseline that replaces fixed vocabulary partitions with greenlists derived from sentence embeddings. That design makes the greenlist context-sensitive. The intervention evaluated here takes a different route: the greenlist remains static, its density is reduced, and the watermark bias is concentrated at higher-entropy positions. Black-box and post-hoc watermarking approaches such as PostMark (Chang et al., 2024) and SimMark (Dabiriaghdam & Wang, 2025) further broaden the deployment space, especially when model-internal access is unavailable. They address a related but distinct problem from the cross-model generation-time evaluation studied here.

Post-hoc AI-generated text detection faces the same basic robustness pressure. Sadasivan et al. (2023) argue that sufficiently strong paraphrasing can remove or obscure detector signals. DetectGPT (Mitchell et al., 2023) relies on probability curvature under the generating model, while RADAR (Hu et al., 2023) uses adversarial training against paraphrasing. Watermarking changes the setup by inserting a keyed signal during generation, but it still needs model-specific calibration, false-positive control, and robustness testing under plausible rewriting.

For information-provenance workflows, calibration is part of the mechanism rather than a reporting afterthought. A detector is difficult to audit if it reports high accuracy without a controlled clean-text false-positive operating point. This motivates reporting attacked-text TPR at an empirical 1% FPR threshold instead of relying only on raw detection accuracy.

Text quality remains a practical constraint for logit-bias watermarking. Aaronson (2022) and Kirchenbauer et al. (2024) note that forcing token probabilities away from the model's natural distribution can impose a fluency or perplexity cost, especially at low-entropy positions where the next token is already predictable. Recent alignment-focused evidence also suggests that watermarking can affect helpfulness and safety beyond perplexity alone (Verma et al., 2026). For that reason, this paper limits its quality claims to the perplexity and attack-validity

measurements actually evaluated.

The remaining gap is model-regime behavior under post-generation rewriting. Prior work establishes important watermark constructions and attack models, but less is known about how greenlist density interacts with each model's effective English vocabulary after paraphrase. This paper addresses that gap by applying the same sparse entropy-gated configuration across four LLM families and by treating BLOOM-7B1's reversal as an informative boundary case rather than as an outlier to discard.

## 3. MATERIALS AND METHODS

### *3.1. Threat Model and Detection Setting*

We evaluate a black-box rewrite setting. The attacker knows that a watermark may be present and may apply paraphrase or backtranslation attacks, but does not know the secret key k or the induced greenlist. An attack output is counted only if it remains a quality-preserving rewrite under the evaluation filters: Sentence-BERT (SBERT) similarity at least 0.70 and attack perplexity at most 300.0. The attacker's objective is to suppress detection by driving the watermark score below the calibrated threshold $\tau$.

The oracle-minimum metric represents a strong single-round attack within the tested attack set. For each prompt, detection is evaluated against the lowest valid post-attack z-score among the T5-PAWS-family rewrite endpoints available for that sample. This is stricter than reporting one paraphraser in isolation, but it is not a claim of robustness against sequential adaptive composition, key disclosure, or a white-box paraphraser that explicitly targets the watermark statistic. The evaluation therefore measures robustness to blind quality-preserving rewriting, not resistance to an attacker with detector queries and adaptive token-level feedback.

The keyed greenlist is central to this audit setting. Without the key, the attacker cannot directly identify green tokens and must instead rely on blind rewriting or statistical inference from queries. Sparsity can reduce the chance that an uninformed token-level replacement falls in the greenlist, since the greenlist occupies fraction $\gamma$ of the vocabulary: 0.25 for AEG rather than 0.50 for KGW. This intuition is not treated as a proof of spoofing resistance, since query volume, threshold calibration, text length, and token-frequency structure also matter. The claim tested here is narrower: under the evaluated black-box rewrite attacks and controlled false-positive thresholds, sparse entropy-gated watermarking retains materially more detection power in several model regimes.

For deployment, the keyed greenlist assignment should be implemented with standard cryptographic key management and a secure pseudorandom function. The experimental comparison assumes that the key remains secret for both AEG and KGW; attacks that recover or brute-force the key are outside the scope of the empirical evaluation.

### *3.2. Baseline Greenlist Watermark*

We use the static UNIGRAM greenlist family studied by Kirchenbauer et al. (2024). Static means that a fixed secret key k induces the same greenlist G for every token position; the preceding context does not change the partition. This differs from context-dependent schemes such as SIR (Zhao et al., 2023). The entropy gate introduced below also preserves this static partition. It changes only whether a bias is applied at a given generation step.

Let V be the model vocabulary and let $\pi_k$ be the keyed pseudorandom permutation used to order its tokens. For greenlist fraction $\gamma$, the greenlist is

$$G = \{ v\_\pi_{k(i)} : 1 \le i \le \lfloor \gamma |V| \rfloor \} \quad (1)$$

The KGW baseline uses $\gamma = 0.50$ with no entropy gate. The sparse AEG configuration uses the same partition family with $\gamma = 0.25$ and adds the entropy-dependent injection rule described in Section 3.4. At generation time, green tokens receive a logit bias $\delta$ whenever injection is active. The detector is intentionally kept identical across KGW and AEG except for the configuration-specific greenlist fraction and calibrated threshold.

Detection counts green tokens in the generated text and evaluates the standard z-score

$$c = \Sigma_{i=1}^{T} 1[t_i \in G] \quad (2)$$
$$z = (c - T\gamma) / \sqrt{(T\gamma(1 - \gamma))}$$

where T is the evaluated token count after context-token skipping and any detection-window truncation. In all main experiments, the detector ignores the first two generated tokens and requires at least 100 surviving tokens. Under an ideal ungated UNIGRAM model, the clean-text null statistic is approximately standard normal. Under AEG, injection is entropy-dependent, so false-positive control is based on empirical clean-text thresholds rather than on the normal approximation alone.

### *3.3. Vocabulary-Regime Design Rule*

The method is motivated by a deployment question: when does making the greenlist sparse help after paraphrasing? The nominal tokenizer size is not enough to answer this question, because an English generation task may use only a subset of a model's full multilingual or byte-level vocabulary. We therefore use the following relation as an empirical design rule rather than as a formal theorem.

Vocabulary-regime design rule. Let the effective English vocabulary be $V_{eff} = |\{v \in V : p_v \ge \varepsilon\}|$, where $\varepsilon$ is a token-frequency floor. This is a conceptual quantity that could be estimated from generation statistics. The absolute greenlist size at greenlist fraction $\gamma$ is $G_{abs}(\gamma) = \lfloor \gamma |V| \rfloor$. We expect sparse AEG ($\gamma = 0.25$) to be most likely to improve post-paraphrase robustness over KGW ($\gamma = 0.50$) when

$$V_{eff} = |\{v \in V : p_v \ge \varepsilon\}| \quad (3)$$
$$G_{abs}(\gamma) = \lfloor \gamma |V| \rfloor$$
$$G_{abs}(0.25) < V_{eff}, \text{ i.e., } 0.25|V| < V_{eff}$$

When this relation holds, the sparse greenlist remains

small relative to the tokens the model actively uses in English, so a paraphraser must remove a more concentrated signal to push z below the calibrated threshold. If the relation fails, or is close to failing, the advantage of lowering γ may disappear: tokenizer-specific frequency structure, calibration behavior, or multilingual subword allocation can dominate the sparse-signal effect. BLOOM-7B1 illustrates this risk. Its nominal vocabulary is the largest in our evaluation, but $\gamma = 0.25$ still produces an approximately 63K-token greenlist; if BLOOM's effective English vocabulary is much smaller than its full tokenizer, KGW can remain stronger. The observed 15.0 percentage-point KGW advantage on BLOOM-7B1 is consistent with this interpretation, but it does not prove the design rule because Veff is not directly estimated here. A direct estimate of Veff from model-specific generation frequencies is left for future work.

*3.4. Adaptive Entropy Gate*

At each generation step, the Shannon entropy of the next-token distribution is computed from the model logits:

$$H_i = -\Sigma_{v \in V} p_i(v) \log_2 p_i(v) \quad (4)$$
$$p_i(v) = \exp(l_i(v)) / \Sigma_{v' \in V} \exp(l_i(v'))$$

where $l_i(v)$ is the logit of token v at position i, and entropy is measured in bits. The gate maintains a rolling buffer Wi containing the most recent 200 entropy values. In the AEG configuration evaluated here, the lower and upper gate thresholds are the 25th and 75th percentiles of this buffer:

$$\hat{q}_{25}(i) = p_{25}(W_i), \quad \hat{q}_{75}(i) = p_{75}(W_i) \quad (5)$$

These rolling percentiles adapt the gate to the entropy profile of each generation without hand-selecting a fixed entropy threshold for each model.

The effective logit bias for token v at step i is

$$\Delta l_i(v) = \min\{\delta \cdot s(H_i; \hat{q}_{25}(i), \hat{q}_{75}(i)), \delta\_cap\} \cdot 1[v \in G] \quad (6)$$

where $\delta\_cap = 3.5$ in the experiments. The gate factor $s \in [0, 1]$ follows a cubic smoothstep:

$$s(H; a, b) = \begin{cases} 0, & \text{if } H \le a \\ 3t^2 - 2t^3, & \text{if } a < H < b,\ t = (H - a)/(b - a) \\ 1, & \text{if } H \ge b \end{cases} \quad (7)$$

Thus positions below the rolling lower quartile receive no watermark bias, positions above the rolling upper quartile receive the full available bias, and intermediate positions receive a smooth partial bias. The realized gate budget is not fixed exactly at one half of positions, because entropy values can be skewed or non-stationary within a generation. We therefore treat the percentile rule as a model-adaptive injection policy, not as an exact analytical budget identity.

The operating intuition is that low-entropy positions are often predictable function tokens or punctuation. Biasing those positions can affect the clean z-score distribution while adding little signal that a paraphraser is likely to preserve or need to remove. AEG instead concentrates watermark pressure on higher-entropy positions, where alternative lexical choices are more plausible. Detection is deliberately kept unchanged: it scores all retained tokens rather than applying an entropy mask. As a result, any gain should be attributed to the combined sparse-gating configuration under empirical threshold calibration, not to a claim that the gate alone causally lowers the clean-text threshold.

***3.5. Threshold Calibration***

False-positive control is empirical throughout the main evaluation. For each model and watermark configuration, we generate a matched clean-text sample and set the reporting threshold τ to the empirical 99th percentile of the clean z-score distribution. The main AEG/KGW results use $N = 300$ clean texts per condition and the lower/order-statistic p99 convention. This gives a common TPR@1%FPR operating point without relying on the standard normal approximation, which may be inaccurate once entropy-dependent injection changes the clean null distribution. These thresholds are analysis thresholds for comparing configurations under the same finite-sample protocol, not deployment thresholds estimated once and reused unchanged across future data.

In deployment, the same rule would be applied to a separate clean calibration set for the target model, key, generation policy, and detector window. Clean and watermarked generation must use matched sampling parameters, since asymmetric sampling changes the clean z-score distribution and can invalidate the false-positive target. In the main runs, temperature is 1.0 and top-k is 50 for all models; Qwen2-7B also uses a repetition penalty of 1.1 in both clean and watermarked paths. The experimental thresholds reported here should therefore be read as controlled evaluation thresholds, while deployment requires a fresh model-specific calibration set.

The watermark strength δ is selected as a quality-constrained operating point rather than treated as a universal optimum. Candidate values are drawn from the fixed grid

$$\delta \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5\}$$

and retained only when they satisfy both quality gates: generating-model self-perplexity at most 25.0 and a watermarked-to-clean self-perplexity ratio at most 1.8. Among retained candidates, the selection score is

$$\text{score}(\delta) = 100 \cdot \text{det rate} + \bar{z}\text{wm} - 10 \max(0, \text{PP}\bar{L}\text{wm} - 22.0)$$

where det rate is the detection rate on the held-out pilot set, $\bar{z}$wm is the mean watermarked z-score, and the final term penalizes mean self-perplexity above 22.0. The selected values are therefore strong enough to create a measurable watermark signal while remaining within the stated quality constraints. This score is used only to choose a quality-constrained operating point; the main conclusions come from the held-out $N = 300$ evaluation rather than from pilot-set optimization performance. The selected operating points are reported in Table 2.

Table 1 summarizes all mathematical symbols used throughout the paper for reference.

**Table 1:** Summary of notation used throughout this paper.

| Symbol | Definition |
|---|---|
| V | Full model vocabulary |
| G | Greenlist: watermarked subset of V |
| k | Secret key for greenlist partition |
| γ | Greenlist fraction = \|G\|/\|V\| |
| V_eff | Effective English vocabulary size; tokens with generation freq. ≥ε, where ε is a token-level frequency floor used in the vocabulary-regime design rule |
| δ | Nominal logit bias applied to green tokens |
| δ_cap | Hard ceiling on effective bias |
| τ | Empirical detection threshold at the target false-positive rate |
| T | Evaluated token count after context-token skipping and truncation |
| z | Watermark z-score statistic |
| $H_i$ | Shannon entropy (bits) of next-token distribution at step i |
| W | Rolling buffer of recent per-token entropy values |
| B | Rolling-buffer size; B = 200 in all experiments |
| $\hat{q}_{25}$, $\hat{q}_{75}$ | Empirical 25th/75th percentiles of W (gate thresholds) |
| s(·) | Cubic smoothstep gate factor ∈ [0, 1] (Eq. 7) |
| N_calib | Number of calibration texts |
| N_test | Number of test texts per condition |

### *3.6. Generation and Detection Workflow*

Algorithm 1 summarizes the AEG generation and detection procedure used in the main experiments. The generator applies the keyed greenlist partition throughout decoding, but the entropy gate controls the strength of the bias at each position. Detection then uses the same greenlist and the empirical threshold τ; it does not require generation logits.

**Algorithm 1.** AEG: Adaptive Entropy-Gated Watermark Generation and Detection.

| **Input: prompt prefix $x_1$:m, model M, secret key k, γ = 0.25, δ, δ_cap = 3.5, buffer size B = 200, and detection threshold τ.** | |
|---|---|
| **Generation** | |
| 1 | G ← GreenList(V, k, γ). |
| 2 | W ← [ ]. |
| 3 | For each generated continuation position i = 1, ..., L: |
| 4 | $l_i$ ← M.logits($x_1$:m, $y_1$:i−1). |
| 5 | $\pi_i$ ← softmax($l_i$). |
| 6 | $H_i \leftarrow -\Sigma_v\, \pi_i(v) \log_2 \pi_i(v)$. |
| 7 | Append $H_i$ to W; if \|W\| > B, drop the oldest value. |
| 8 | a ← $p_{25}$(W); b ← $p_{75}$(W). |
| 9 | If b ≤ a, set b ← a + ε. |
| 10 | t ← clamp(($H_i$ − a)/(b − a), 0, 1). |
| 11 | gate ← $3t^2 - 2t^3$. |
| 12 | δ_eff ← min(δ · gate, δ_cap). |
| 13 | For all v ∈ G, set $l_i(v) \leftarrow l_i(v) + δ\_eff$. |
| 14 | Sample $y_i \sim$ softmax($l_i$). |
| **Detection** | |
| 15 | Tokenize the generated continuation $y_1$:L to ids ($u_1$, ..., u_L). |
| 16 | Discard the first two generated tokens and keep at most 512 evaluated tokens. |
| 17 | T ← number of retained evaluated tokens. |
| 18 | If T < 100, return insufficient length. |
| 19 | $c \leftarrow \Sigma_{j=1}^{T} 1[u_j \in G]$. |
| 20 | $z \leftarrow (c - Tγ) / \sqrt{(Tγ(1 - γ))}$. |
| 21 | Return watermarked if z > τ; otherwise return clean. |

### *3.7. Experimental Setup*

The main evaluation uses N = 300 generated texts per model and condition, with matched clean samples for empirical thresholding. Prompts are drawn from a heterogeneous pool of 148 custom prompts plus 360

WikiText-2 passage openings, yielding 508 prompts in total (Merity et al., 2016). We evaluate four open-weight 7-8B-parameter LLMs with different tokenizer regimes: LLaMA-3-8B (128K tokens) (Meta AI, 2024), Mistral-7B-v0.1 (32K tokens) (Jiang et al., 2023), Qwen2-7B (152K tokens) (Qwen Team, 2024), and BLOOM-7B1 (251K tokens) (BigScience Workshop, 2022).

The primary comparison is between KGW, implemented as a static UNIGRAM greenlist watermark with $\gamma = 0.50$ and no entropy gate, and AEG, which uses the same greenlist family with $\gamma = 0.25$ and entropy-gated injection. We also retain a constrained SIR row as contextual evidence about context-dependent greenlisting, but not as the central baseline for the paper. All main AEG/KGW runs use bitsandbytes 4-bit NF4 quantization (Dettmers et al., 2023). Generation uses temperature 1.0, top-k = 50, at least 250 new tokens, and at most 600 new tokens; Qwen2-7B uses a matched repetition penalty of 1.1 in both clean and watermarked generation. The main runs were carried out in dual 16 GB NVIDIA T4 Kaggle notebook environments. The supplementary evidence bundle preserves result arrays, configuration summaries, selected analysis scripts, environment notes, and a file manifest sufficient to check the reported aggregate values. Complete environment details are preserved for the main N = 300 AEG/KGW comparison; auxiliary key-sensitivity and calibration runs are therefore treated as deployment diagnostics rather than replacements for the main evaluation.

We evaluate robustness against three T5-PAWS paraphrase variants (Raffel et al., 2020; Zhang et al., 2019), denoted Weak, Strong, and Recursive, and against English–French–English backtranslation using Helsinki-NLP OPUS-MT (Tiedemann & Thottingal, 2020). Attack outputs are accepted only when they pass the quality filters described below. The headline robustness metric is oracle-minimum TPR@1%FPR: for each sample, we use the lowest valid post-attack z-score among the T5-PAWS-family rewrite endpoints preserved for that sample. This is a strong single-round stress test within the evaluated paraphrase family, while sequential adaptive composition remains outside the tested scope.

Table 2 summarizes the model configurations and selected watermark strengths.

**Table 2:** Model configurations and selected logit-bias values. Quant. = bitsandbytes 4-bit NF4. $\delta$ denotes the quality-constrained operating point selected by the procedure in Section 3.5. SIR is included only as a contextual schema comparison.

| Model | Vocab | Quant. | $\delta$ | | |
|---|---|---|---|---|---|
| | | | AEG | KGW | SIR |
| LLaMA-3-8B | 128K | 4-bit NF4 | 3.5 | 3.5 | 3.5 |
| Mistral-7B | 32K | 4-bit NF4 | 3.5 | 2.5 | 3.5 |
| Qwen2-7B | 152K | 4-bit NF4 | 3.0 | 1.0 | 1.0[a] |
| BLOOM-7B1 | 251K | 4-bit NF4 | 3.5 | 3.5 | 3.5 |

### *3.8. Prompt Dataset*

The prompt pool combines 148 custom prompts with 360 WikiText-2 passage openings, yielding 508 prompts in total (Merity et al., 2016). Prompts range from 5 to 80 words and cover varied informational, explanatory, and argumentative writing tasks. We partition the pool into non-overlapping calibration, pilot, seed, and test subsets with a fixed random seed, so no prompt appears in more than one role. The supplementary bundle includes a prompt source and split manifest describing the source counts, shuffle seed, split sizes, and redistribution limits for the prompt pool.

### *3.9. Rewrite Attacks and Validity Filters*

Attack outputs are counted only when they remain quality-preserving under two automatic filters. First, SBERT cosine similarity between the pre-attack generated text and the corresponding attacked text must be at least 0.70. Second, perplexity under the generating model must not exceed 300.0. Similarity is computed with the all-MiniLM-L6-v2 Sentence-Transformers model (Reimers & Gurevych, 2019). This policy prevents meaning-destroying rewrites from being treated as successful evasions; post-attack TPRs are therefore conditional on attacks that remain plausible paraphrases. The filters remove obvious attack failures, but they do not replace human semantic evaluation.

Across the four model families, AEG and KGW T5-PAWS validity is 67%-86%, and backtranslation validity is 96%-99%. Accepted attacks remain semantically close to the source text. For AEG, the count-weighted mean SBERT similarity is 0.823 for Weak, 0.818 for Strong, 0.819 for Recursive, and 0.928 for backtranslation.

The accepted-attack denominators also do not explain the main Qwen2-7B result. KGW has more accepted T5-PAWS attacks than AEG on Weak (245/300 vs. 216/300), Strong (222/300 vs. 202/300), and Recursive (226/300 vs. 215/300), yet AEG remains far more detectable after attack. Table 3 reports the per-model validity rates and accepted-attack SBERT means.

**Table 3:** Attack validity rates and mean SBERT similarity. Validity requires SBERT ≥ 0.70 and attack-PPL ≤ 300.0. Means are computed over accepted attacks only. BT = EN→FR→EN backtranslation via Helsinki-NLP.

| Model | Method | Metric | Weak | Strong | Recur. | BT |
|---|---|---|---|---|---|---|
| LLaMA-3-8B | AEG | Valid (%) | 85 | 77 | 79 | 97 |
| | | SBERT | 0.828 | 0.822 | 0.823 | 0.932 |
| | KGW | Valid (%) | 84 | 79 | 85 | 99 |
| | | SBERT | 0.833 | 0.823 | 0.828 | 0.927 |
| Mistral-7B | AEG | Valid (%) | 80 | 72 | 78 | 97 |
| | | SBERT | 0.825 | 0.820 | 0.825 | 0.925 |
| | KGW | Valid (%) | 78 | 72 | 80 | 99 |
| | | SBERT | 0.815 | 0.807 | 0.819 | 0.922 |
| Qwen2-7B | AEG | Valid (%) | 72 | 67 | 72 | 99 |
| | | SBERT | 0.804 | 0.801 | 0.804 | 0.921 |
| | KGW | Valid (%) | 82 | 74 | 75 | 98 |
| | | SBERT | 0.810 | 0.809 | 0.806 | 0.916 |
| BLOOM-7B1 | AEG | Valid (%) | 86 | 79 | 83 | 97 |
| | | SBERT | 0.831 | 0.827 | 0.821 | 0.933 |
| | KGW | Valid (%) | 86 | 82 | 84 | 96 |
| | | SBERT | 0.822 | 0.815 | 0.813 | 0.931 |

### *3.10. Metrics and Statistical Tests*

All headline TPR values use the empirical threshold $\tau$ derived from the corresponding N = 300 clean z-score distribution, as described in Section 3.5. Attack TPRs are computed over quality-passing attack outputs, and Wilson intervals summarize binomial uncertainty. Paired AEG–KGW evidence uses continuity-corrected McNemar tests on paired oracle-minimum detection outcomes. We use a family-wise significance level of $\alpha = 0.05$ for this planned family. The Bonferroni correction uses m = 12 because paired tests were pre-specified across three T5-PAWS attack endpoints and four models. Table 6 reports the oracle-minimum endpoint as a compact summary. For Mistral-7B, BLOOM-7B1, and Qwen2-7B, Table 6 uses the deposited master paired-analysis output files for an independent T5-PAWS-only paired evaluation. The LLaMA-3-8B paired result is retained only as aggregate diagnostic output because its full paired arrays are unavailable; the main LLaMA-3-8B conclusion therefore rests on operating-point TPRs, confidence intervals, and margin diagnostics rather than independently deposited paired arrays.

## 4. RESULTS

### *4.1. Main TPR at 1% FPR Results*

Table 4 reports TPR@1%FPR for KGW, AEG, a gate-on $\gamma = 0.50$ ablation, and a constrained SIR schema-comparison row. Because SIR uses a different schema and a smaller constrained run, it is not used to define the headline AEG/KGW claim. The main pattern is that AEG improves post-paraphrase detection in three of the four model families. It leads KGW on LLaMA-3-8B, Mistral-7B, and Qwen2-7B under the Weak, Strong, Recursive, and oracle-minimum attacks. The largest provenance-detection gain appears on Qwen2-7B: weak-attack TPR rises from 38.4% to 99.5%, and oracle-minimum TPR rises from 17.3% to 98.8% under the same quality filters and empirical false-positive operating point.

The result does not follow a monotonic tokenizer-size pattern. BLOOM-7B1 has the largest nominal vocabulary in the study, yet it is the one regime where KGW remains stronger: KGW reaches 87.4% oracle-minimum TPR compared with 72.4% for AEG. This reversal supports the paper's main interpretation that watermark robustness depends on the relationship between greenlist density, empirical clean-score calibration, and the model's actively used English vocabulary, rather than on nominal vocabulary size alone.

The ablation row also shows that the entropy gate is not sufficient by itself. With the gate enabled but $\gamma = 0.50$, oracle-minimum TPR falls to 3.7% on LLaMA-3-8B, 71.6% on Mistral-7B, 15.4% on Qwen2-7B, and 17.1% on BLOOM-7B1. In the tested design space, the robust configuration is therefore the sparse gated setting, not a dense greenlist with gating added afterwards.

Backtranslation is less discriminating on LLaMA-3-8B and Mistral-7B, where both methods remain high and KGW is ahead by only 0.1 and 0.7 percentage points. Qwen2-7B is different: AEG reaches 99.7% backtranslation TPR, while KGW falls to 67.2%. Table 5 gives Wilson 95% confidence intervals. The AEG and KGW oracle-minimum intervals do not overlap on LLaMA-3-8B or Qwen2-7B. On Mistral-7B they meet at the boundary ([94,98] for AEG and [87,94] for KGW), so the gain should be read as positive but smaller. BLOOM-7B1 shows the reverse pattern, with disjoint intervals favoring KGW.

Table 6 summarizes paired evidence for the AEG–KGW comparison, and Table 7 reports descriptive weak-attack margins at the main empirical thresholds. McNemar tests on paired oracle-minimum detection outcomes support AEG on Mistral-7B and Qwen2-7B from deposited paired arrays; the LLaMA-3-8B row is shown as aggregate diagnostic support only. BLOOM-7B1 is not significant and remains a KGW-favored regime in the main result

table. These paired tests do not replace the main TPR estimates; they show that the observed direction of effect is not merely a by-product of unpaired confidence intervals.

Figure 1 visualizes the weak and oracle-minimum operating points. Figure 2 shows the AEG clean, watermarked, and weak-attack z-score separation used for empirical thresholding, and Figure 3 summarizes the AEG–KGW ordering across a wider FPR range.

**Table 4.** TPR@1%FPR (%) across four LLMs. AEG is the sparse entropy-gated configuration (γ = 0.25); KGW is the greenlist baseline (γ = 0.50, no gate). AEGγ = 0.5 tests gating without sparsity. SIR is a constrained schema-comparison row with N = 100. Bold marks the better AEG/KGW result per column. Oracle-minimum uses the lowest valid post-attack z-score among the T5-PAWS-family rewrite endpoints.

| Model | Method | γ | WM | Weak | Strong | Recur. | BT | Oracle-min.[b] |
|---|---|---|---|---|---|---|---|---|
| LLaMA-3-8B (128K) | KGW (Kirchenbauer et al., 2023) | 0.50 | 98.3 | 69.7 | 61.6 | 61.3 | 93.9 | 53.3 |
| AEG (ours) | 0.25 | 96.7 | 93.3 | 92.2 | 91.2 | 93.8 | 88.7 | |
| $AEG_{\gamma=0.5}$ | 0.50 | 94.7 | 14.0 | 10.5 | 8.3 | 78.5 | 3.7 | |
| SIR (Zhao et al., 2023)[a] | 0.25[a] | 96.0 | 1.3 | 1.5 | 0.0 | 70.0 | 0.0 | |
| Mistral-7B (32K) | KGW (Kirchenbauer et al., 2023) | 0.50 | 100.0 | 96.2 | 94.4 | 94.1 | 100.0 | 91.4 |
| AEG (ours) | 0.25 | 100.0 | 98.8 | 97.7 | 98.3 | 99.3 | 96.9 | |
| $AEG_{\gamma=0.5}$ | 0.50 | 100.0 | 86.5 | 77.1 | 81.3 | 99.7 | 71.6 | |
| SIR (Zhao et al., 2023)[a] | 0.25[a] | 98.0 | 18.7 | 10.8 | 19.0 | 87.0 | 2.2 | |
| Qwen2-7B (152K) | KGW (Kirchenbauer et al., 2023) | 0.50 | 90.7 | 38.4 | 32.0 | 28.8 | 67.2 | 17.3 |
| AEG (ours) | 0.25 | 100.0 | 99.5 | 99.0 | 99.5 | 99.7 | 98.8 | |
| $AEG_{\gamma=0.5}$ | 0.50 | 92.7 | 27.9 | 18.5 | 25.7 | 71.4 | 15.4 | |
| SIR (Zhao et al., 2023)[a] | 0.25[a] | 68.0 | 9.9 | 5.5 | 5.1 | 23.5 | 1.2 | |
| BLOOM-7B1 (251K) | KGW (Kirchenbauer et al., 2023) | 0.50 | 100.0 | 91.4 | 92.7 | 90.0 | 100.0 | 87.4 |
| AEG (ours) | 0.25 | 100.0 | 88.7 | 81.4 | 77.1 | 99.0 | 72.4 | |
| $AEG_{\gamma=0.5}$ | 0.50 | 98.0 | 35.9 | 33.1 | 25.6 | 95.2 | 17.1 | |
| SIR (Zhao et al., 2023)[a] | 0.25[a] | 100.0 | 15.0 | 9.1 | 9.2 | 93.7 | 2.5 | |

**Table 5.** TPR@1%FPR (%) with Wilson 95% confidence intervals. Format: point estimate [lo, hi]. AEG/KGW rows use up to N = 300 generated texts per condition; SIR is a smaller constrained schema-comparison run. SIR attack cells with valid-attack N < 30 are shown as point estimates only. TPR@1%FPR uses the empirical p99 clean-score threshold, so the realized finite-sample clean-tail rate is approximate rather than exactly continuous 1%.

| Model | Method | WM | Weak | Strong | Recur. | BT | Oracle-min.[c] |
|---|---|---|---|---|---|---|---|
| LLaMA-3-8B | AEG | 96.7 [94,98] | 93.3 [90,96] | 92.2 [88,95] | 91.2 [87,94] | 93.8 [90,96] | 88.7 [84,92] |
| KGW | 98.3 [96,99] | 69.7 [64,75] | 61.6 [55,68] | 61.3 [55,67] | 93.9 [91,96] | 53.3 [47,59] | |
| $AEG_{\gamma=0.5}$ | 94.7 [92,97] | 14.0 [10,19] | 10.5 [7,15] | 8.3 [5,12] | 78.5 [74,83] | 3.7 [2,7] | |
| SIR[a],b | 96.0 [90,98] | 1.3 [0,7] | 1.5 [0,8] | 0.0 [0,5] | 70.0 [60,78] | 0.0 [0,4] | |
| Mistral-7B | AEG | 100.0 [99,100] | 98.8 [96,100] | 97.7 [95,99] | 98.3 [96,99] | 99.3 [98,100] | 96.9 [94,98] |
| KGW | 100.0 | 96.2 | 94.4 | 94.1 | 100.0 | 91.4 | |

| Model | Method | WM | Weak | Strong | Recur. | BT | Oracle-min.[c] |
|---|---|---|---|---|---|---|---|
| | [99,100] | [93,98] | [90,97] | [90,96] | [99,100] | [87,94] | |
| $AEG_{\gamma=0.5}$ | 100.0 [99,100] | 86.5 [82,90] | 77.1 [71,82] | 81.3 [76,86] | 99.7 [98,100] | 71.6 [66,77] | |
| SIR[a],b | 98.0 [93,99] | 18.7 [11,29] | 10.8 [5,21] | 19.0 [12,29] | 87.0 [79,92] | 2.2 [1,8] | |
| Qwen2-7B | AEG | 100.0 [99,100] | 99.5 [97,100] | 99.0 [96,100] | 99.5 [97,100] | 99.7 [98,100] | 98.8 [96,100] |
| KGW | 90.7 [87,93] | 38.4 [33,45] | 32.0 [26,38] | 28.8 [23,35] | 67.2 [62,72] | 17.3 [13,22] | |
| $AEG_{\gamma=0.5}$ | 92.7 [89,95] | 27.9 [22,34] | 18.5 [14,24] | 25.7 [20,32] | 71.4 [66,76] | 15.4 [11,20] | |
| SIR[a],b | 68.0 [58,76] | 9.9 [5,18] | 5.5 [2,13] | 5.1 [2,12] | 23.5 [16,33] | 1.2 [0,6] | |
| BLOOM-7B1 | AEG | 100.0 [99,100] | 88.7 [84,92] | 81.4 [76,86] | 77.1 [72,82] | 99.0 [97,100] | 72.4 [67,77] |
| KGW | 100.0 [99,100] | 91.4 [87,94] | 92.7 [89,95] | 90.0 [86,93] | 100.0 [99,100] | 87.4 [83,91] | |
| $AEG_{\gamma=0.5}$ | 98.0 [96,99] | 35.9 [30,42] | 33.1 [27,39] | 25.6 [20,32] | 95.2 [92,97] | 17.1 [13,22] | |
| SIR[a],b | 100.0 [96,100] | 15.0 [8,26] | 9.1 [4,20] | 9.2 [4,19] | 93.7 [87,97] | 2.5 [1,9] | |

**Table 6.** Paired McNemar evidence for AEG vs. KGW. The test uses paired oracle-minimum detection outcomes with continuity-corrected $\chi^2(1)$. Bonferroni correction uses a planned family-wise $\alpha = 0.05$ over $m = 12$ pre-specified paired tests across three endpoints and four models, giving threshold $p < 4.17\times10^{-3}$.

| Model | b | c | $\chi^2(1)$ | p | Bonf./status | Direction |
|---|---|---|---|---|---|---|
| LLaMA-3-8B[a] | 100 | 10 | 72.01 | $<2\times10^{-16}$ | diag. | AEG |
| Mistral-7B | 84 | 30 | 24.64 | $6.9\times10^{-7}$ | ✓ AEG | |
| BLOOM-7B1 | 13 | 8 | 0.76 | 0.383 | × n.s. | n.s.; KGW |
| Qwen2-7B | 33 | 2 | 25.71 | $4.0\times10^{-7}$ | ✓ AEG | |

**Table 7.** Weak-attack margin diagnostics for AEG vs. KGW at the main empirical p99 thresholds. Margins are the mean of (z-τ) over valid weak-attack outputs in the main N = 300 runs. Positive values indicate that the average valid attacked output remains above the reporting threshold.

| Model | AEG valid N | AEG margin | KGW valid N | KGW margin |
|---|---|---|---|---|
| LLaMA-3-8B | 254 | +3.29 | 251 | +0.62 |
| Mistral-7B | 241 | +4.30 | 235 | +2.35 |
| Qwen2-7B | 216 | +3.88 | 245 | -0.50 |
| BLOOM-7B1 | 257 | +1.90 | 257 | +2.56 |

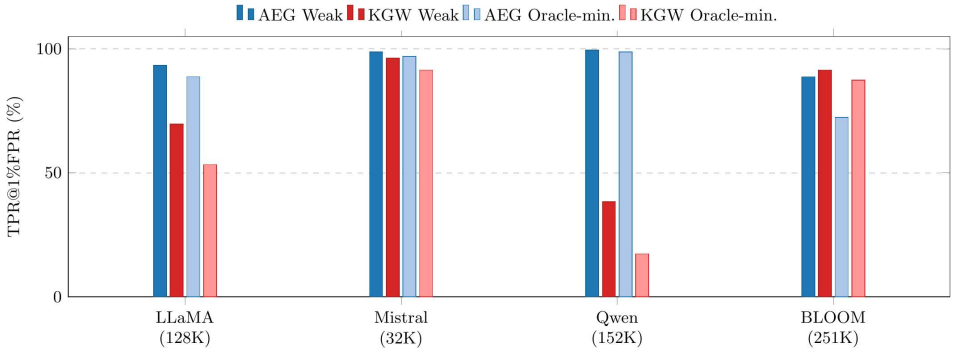


**Figure 1.** Weak and oracle-minimum TPR@1%FPR for AEG and KGW. AEG improves both measures on LLaMA-3-8B, Mistral-7B, and Qwen2-7B. The largest gap is on Qwen2-7B, where oracle-minimum TPR improves by 81.5 percentage points; BLOOM-7B1 marks the KGW-favored regime. Values match the operating-point estimates in Table 4.

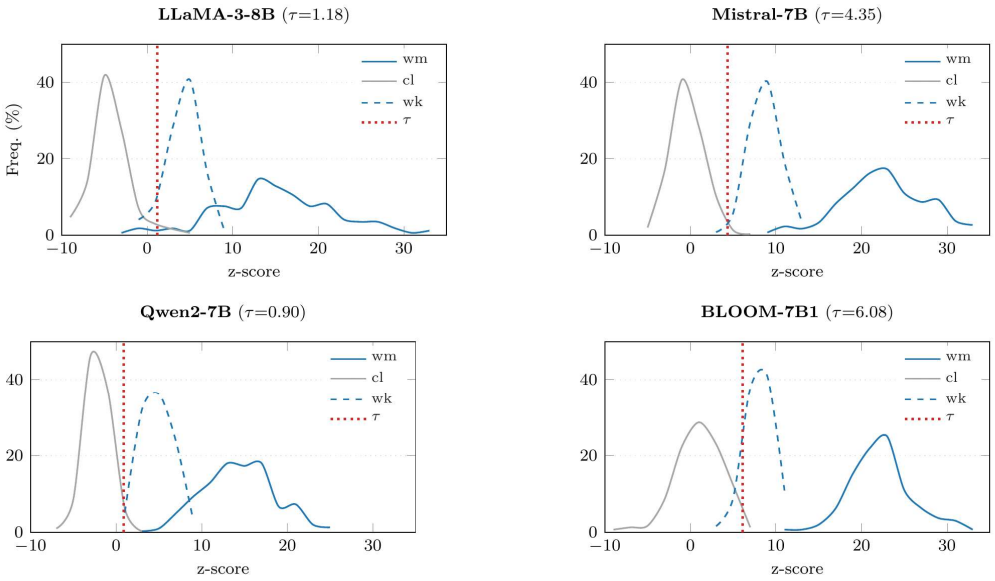


**Figure 2.** AEG z-score summaries for watermarked, clean, and weak-attacked text (N = 300 per model and condition). Solid blue shows watermarked text, gray shows clean text, dashed blue shows weak attacks, and the dotted red line marks the empirical p99 clean-score threshold used for the operating-point TPRs. The panels show why empirical clean-score calibration matters: low clean thresholds leave more post-attack margin on LLaMA-3-8B and Qwen2-7B, while BLOOM-7B1 has a higher threshold and less AEG post-attack separation.

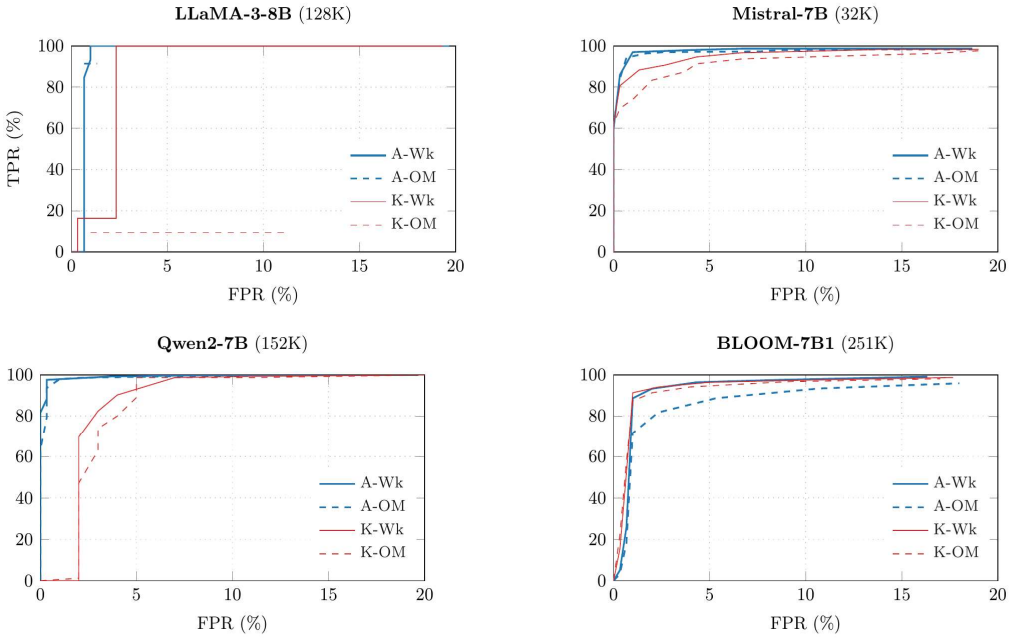


**Figure 3.** ROC-style summaries over FPR 0–20% for AEG (blue) and KGW (red). Solid curves show weak attacks, and dashed curves show oracle-minimum attacks. Clean-side FPR is computed from the N = 300 clean-score distribution, while attack-side TPR uses validity-filtered attack outputs. These curves are included to show relative ordering across operating points; the exact 1% FPR claims are the values reported in Tables 4 and 5.

### *4.2. Ablation Study*

Table 8 separates two design choices: reducing the greenlist fraction to $\gamma = 0.25$ and applying the entropy gate. The ablation is best read as a practical configuration test, not a clean causal isolation. The gate-on $\gamma = 0.50$ row asks whether gating alone is enough; the gate-off $\gamma = 0.25$

row asks whether sparsity alone is enough.

Neither ablation reproduces the full AEG result. With the gate enabled but γ kept at 0.50, oracle-minimum TPR falls to 3.7% on LLaMA-3-8B, 71.6% on Mistral-7B, 15.4% on Qwen2-7B, and 17.1% on BLOOM-7B1. This row underperforms both KGW and full AEG on every model, so the entropy gate alone does not explain the headline gains. Conversely, γ = 0.25 without the gate also fails to match full AEG: oracle-minimum TPR is only 8.0% on LLaMA-3-8B, 49.7% on Mistral-7B, 18.2% on Qwen2-7B, and 20.3% on BLOOM-7B1.

The gate-off rows provide a practical but not fully matched configuration contrast, because backtranslation was unavailable in those runs. They support the qualitative conclusion that sparsity alone is insufficient, but the exact recovery magnitudes should not be interpreted as additive causal estimates. Relative to the gate-off sparse row, oracle-minimum TPR rises by +80.7pp on LLaMA-3-8B, +47.2pp on Mistral-7B, +80.6pp on Qwen2-7B, and +52.1pp on BLOOM-7B1. The same pattern appears under weak attack, with gains of +75.3, +14.4, +64.7, and +45.8pp. These are large configuration-level effects, but they should be interpreted as evidence for the combined sparse entropy-gated configuration rather than as an exact additive gate effect.

The model-specific pattern is informative. On Mistral-7B, γ = 0.25 without the gate still gives a strong pre-attack watermark signal, but it raises the clean-side threshold to τ=7.01, compared with τ=4.35 for full AEG and τ=4.50 for KGW. The gate therefore appears to recover robustness partly by improving the clean-score operating point, not only by moving watermark signal to high-entropy positions. On Qwen2-7B, sparsity alone is almost neutral relative to KGW on oracle-minimum TPR (18.2% vs. 17.3%), but the gate lifts the sparse configuration to 98.8%. BLOOM-7B1 again marks the boundary of the regime: the gate recovers substantial performance (20.3% → 72.4%), but full AEG still remains below KGW (87.4%), consistent with the vocabulary-regime design rule.

**Table 8.** Configuration contrasts for greenlist sparsity and entropy gating. Values are TPR@1%FPR (%) for Weak (Wk) and oracle-minimum (OM[a]). The Gate ON, γ = 0.50 row tests gating without sparse greenlists; the Gate OFF, γ = 0.25 row tests sparsity without gating. Full AEG combines both choices. Gate-OFF OM is T5-PAWS-only because backtranslation was unavailable in those runs; see table note.

| Configuration | LLaMA-3-8B | | Mistral-7B | | Qwen2-7B | | BLOOM-7B1 | |
|---|---|---|---|---|---|---|---|---|
| | Wk | OM[a] | Wk | OM[a] | Wk | OM[a] | Wk | OM[a] |
| KGW (γ = 0.50, no gate) | 69.7 | 53.3 | 96.2 | 91.4 | 38.4 | 17.3 | 91.4 | 87.4 |
| Gate ON, γ = 0.50 | 14.0 | 3.7 | 86.5 | 71.6 | 27.9 | 15.4 | 35.9 | 17.1 |
| Gate OFF, γ = 0.25 | 18.0[b] [14, 23] | 8.0[b] [5, 12] | 84.4[b] [79, 88] | 49.7[b] [44, 55] | 34.8[b] [29, 41] | 18.2[b] [14, 24] | 42.9[b] [36, 50] | 20.3[b] [16, 26] |
| AEG: gate + γ = 0.25 | 93.3 | 88.7 | 98.8 | 96.9 | 99.5 | 98.8 | 88.7 | 72.4 |

## 5. DISCUSSION

The central operating-point result is not only a higher true-positive rate. At the fixed 1% FPR threshold, AEG also changes the detector margin left after rewriting. On LLaMA-3-8B, for example, the weak-attack mean is lower under AEG than under KGW, but AEG is evaluated against a much lower clean-text threshold: AEG sits about 3.29 z-units above its threshold after weak paraphrasing, while KGW sits only about 0.62 units above its threshold. Qwen2-7B shows the same pattern more sharply. AEG remains about 3.88 z-units above threshold after the weak attack, whereas KGW falls about 0.50 units below threshold. These margins explain why the TPR gap is largest on Qwen2-7B even though the nominal change in γ is simple.

The ablations reinforce that the method should be read as a combined sparse entropy-gated configuration, not as an isolated gate effect. Gating with γ = 0.50 is not enough: oracle-minimum TPR falls to 3.7%, 71.6%, 15.4%, and 17.1% on LLaMA-3-8B, Mistral-7B, Qwen2-7B, and BLOOM-7B1. Sparsity without the gate is also insufficient in the gate-off ablation, where oracle-minimum TPR is 8.0%, 49.7%, 18.2%, and 20.3% under the T5-PAWS-only suite. The full AEG configuration is therefore best understood as an operating regime: sparse greenlists can increase the post-attack score margin under the calibrated detector when the model's active token regime supports the sparse configuration, while entropy gating concentrates bias where generation has enough lexical freedom to carry the watermark without excessive quality cost. Because detection is not entropy-masked, these experiments do not prove that the gate alone suppresses the clean null distribution; they show that the combined configuration produces the robust margins after rewriting.

The model ordering supports a vocabulary-regime interpretation and argues against a simple tokenizer-size account. AEG improves oracle-minimum TPR by 35.4 percentage points on LLaMA-3-8B and by 81.5 points on Qwen2-7B, while Mistral-7B shows a smaller 5.5-point gain. BLOOM-7B1 reverses the pattern: KGW reaches 87.4% oracle-minimum TPR compared with 72.4% for AEG. This is the key counterexample. BLOOM has the

largest nominal tokenizer in the study, but its multilingual byte-level BPE vocabulary does not imply the same English-active token regime as Qwen2-7B or LLaMA-3-8B. The result therefore argues against choosing $\gamma$ from tokenizer size alone. What matters in practice is the density of the greenlist within the model's actively used English vocabulary, together with the clean-score threshold induced by that choice.

The SIR results are included as a constrained schema comparison, not as a definitive claim about production SIR. The evaluated SIR variant uses the available continuous-projection adaptation under the same model family and attack protocol, and the near-zero T5-PAWS robustness should be read with that limitation in mind. Its role in the paper is mainly contextual: under this evaluation harness, the sparse entropy-gated configuration provides a stronger provenance signal after local paraphrase attacks than the tested SIR adaptation. A production-level SIR comparison remains outside the present scope because it would require the original LSH-based implementation, matched generation settings, and the same empirical thresholding protocol.

Key sensitivity remains a deployment issue, but it does not overturn the main result. The Qwen2-7B gate-on 10-key sweep shows stable pre-attack injection and lower threshold variation than LLaMA-3-8B: Qwen2-7B gate-on thresholds range from 0.393 to 4.210 with mean 2.59±1.10 and bootstrap CI [1.89, 3.24], with no key above the run-specific screening cutoff of $\tau$=4.78. LLaMA-3-8B is less stable: excluding one anomalous low-threshold key, its gate-on mean is 4.601±1.987 with CI [3.39, 5.99], and three of nine non-anomalous keys exceed the corresponding $\tau$=4.46 screening cutoff. These sweeps support threshold-stability screening and key selection before deployment; they should not be used as direct replacements for the main N = 300 attack evaluation, because their calibration budget and attack suite differ.

These results support a calibrated provenance-detection claim. The work does not propose a new cryptographic primitive; it shows that, under controlled false-positive calibration, sparse entropy-gated watermarking can materially improve post-paraphrase detection in three model regimes while failing to dominate in a fourth. The supplementary diagnostics report the calibration, key-sensitivity, z-score, and perplexity details separately to support reproducibility.

The practical implication is procedural. When LLM-generated text may later be revised, translated, moderated, archived, or published, a watermark detector should not be reported as a standalone score. It should be accompanied by the calibration protocol, operating threshold, attack suite, model/key regime, validity filters, and enough supporting evidence for audit. This framing keeps AEG as a calibrated provenance component rather than a complete governance or security system.

### *5.1. Limitations*

The evaluation is limited to sentence-level and round-trip rewrite attacks. T5-PAWS substitutes local sentence content, and Helsinki-NLP backtranslation preserves broad sentence order. The oracle-minimum metric tests these attacks independently, not as a chained rewrite process. Stronger discourse-level rewriting, repeated adaptive paraphrasing, and instruction-tuned LLM paraphrasers could alter green-token structure more aggressively. Two gate-aware attacks also remain unevaluated: prompt crafting that suppresses entropy-gate activation and paraphrasers that preferentially rewrite high-entropy positions.

The ablation evidence is configuration-level, not a clean causal decomposition. The gate-on $\gamma$ = 0.50 row and the gate-off $\gamma$ = 0.25 row show that neither gating alone nor sparsity alone explains the main gains, but the gate-off oracle-minimum values are T5-PAWS-only because backtranslation was unavailable. The LLaMA-3-8B gate-off run also used a different detection window, so its raw threshold should not be compared directly with the main LLaMA threshold. These caveats do not change the main AEG–KGW comparison in Tables 4 and 5; they limit only how far the ablation can be interpreted as an additive gate effect.

The reported TPR values use empirical p99 thresholds estimated from held-out clean test scores. This is appropriate for controlled comparison at 1% FPR, but it is not yet a deployment recipe. Earlier small-sample calibration trials with N_calib = 100 were unstable in several conditions, so a deployed detector should use a larger clean calibration set, key screening, and a standardized detection window. The present implementation also uses a reproducibility-oriented greenlist seed rather than a cryptographically secure key derivation procedure; deployment would require replacing it with a CSPRNG/PRF-based construction.

The vocabulary-regime explanation is supported by four models, but it is not a direct measurement of each model's effective English vocabulary. BLOOM-7B1 is an important counterexample to nominal tokenizer-size reasoning, yet its multilingual byte-level tokenizer also introduces a confound: the English-active vocabulary may be much smaller than the full tokenizer. The study is English-only, so cross-lingual watermark survival is outside the scope of the present evidence.

Finally, text quality is evaluated through perplexity filters and SBERT-based attack validity, without human assessment. These automatic checks are useful for controlling obvious failures, but they do not measure human fluency, factuality, helpfulness, safety, or whether accepted paraphrases remain useful in downstream tasks. The experiments also use 4-bit NF4 quantization for feasibility on commodity hardware. A full-precision reproduction could reveal small model-dependent changes in gate activation, clean-score variance, or margins after rewriting. Taken together, these limits mean that the paper provides a calibrated comparative evaluation, not a deployment guarantee.

## 6. CONCLUSIONS

Sparse entropy-gated watermarking is most useful here as a calibrated information-provenance configuration, not as a universal replacement for KGW. Under empirical 1% FPR calibration, AEG improves oracle-minimum post-attack TPR over KGW on LLaMA-3-8B, Mistral-7B, and Qwen2-7B. The Qwen2-7B gain is the clearest: TPR rises from 17.3% to 98.8%; LLaMA-3-8B also rises from 53.3% to 88.7%. Mistral-7B shows a smaller positive gap, while BLOOM-7B1 favors KGW. That negative case is essential because it separates watermark robustness from nominal tokenizer size. When generated text passes through multiple processing stages in a distributed serving or publishing pipeline, calibrated provenance signals let downstream auditors trace its transformations directly.

The evidence suggests that $\gamma$ should be treated as a model-regime parameter. Sparse greenlists help when they are sparse relative to the model's actively used English vocabulary and when clean-score calibration leaves enough margin after rewriting. Entropy gating makes that sparse choice usable by placing bias where token choice is less constrained. The resulting claim is empirical and provenance-oriented: robustness after paraphrase depends on greenlist density, entropy-gated injection, clean-text thresholds, and vocabulary regime acting together.

For deployment-oriented evaluation, a watermark should not transfer a single KGW-style setting from one model to another. It should report the false-positive target, attacked-text TPR, calibration sample size, key-screening behavior, and the model regime in which the watermark was tuned. BLOOM-7B1 shows why this validation has to remain model aware. The next evaluation priority is a stronger attack suite, including repeated paraphrasing, discourse-level restructuring, and instruction-tuned LLM paraphrasers. Direct measurement of each model's English-active vocabulary is also needed, so that $\gamma$ can be selected from the operating regime rather than from nominal tokenizer size. Before any deployment use, calibration should be tightened through larger clean calibration sets, prompt-paired key sweeps across all models, harmonized detection windows, and a cryptographically secure greenlist keying procedure. The SIR comparison should also be repeated with the original LSH-based implementation under matched generation and thresholding settings.

**Acknowledgments**

No non-author acknowledgments are reported.

**Ethics Statement**

This study did not involve human participants, human tissue, or personal data.

**AI-Assisted Writing Disclosure**

AI-assisted tools, including ChatGPT-based and Codex-based tools, were used during manuscript preparation for language review, consistency checking, formatting guidance, document-preparation troubleshooting, and LaTeX/package troubleshooting. No AI-assisted tool was used to generate, fabricate, modify, synthesize, or analyze the underlying experimental data. The authors manually reviewed the manuscript, checked the numerical claims against the available evidence files, and take responsibility for the submitted content.

**Funding**

This research received no external funding.

**Author Contributions**

A.U. designed the study, implemented the watermarking pipeline, ran the experiments, analyzed the results, curated the data, prepared the original manuscript draft, and produced the visualizations. V.K.T. and P.K. contributed resources, supported validation, supervised the work, and reviewed and edited the manuscript. All authors have read and approved the submitted version.

**Data Availability and Supplementary Materials**

The processed score arrays, clean-score thresholds, attack-validity summaries, ablation summaries, selected paired-test outputs, selected analysis scripts, evidence map, prompt source and split manifest, environment note, and file manifest are submitted as Supplementary File S1 (supplementary PDF) and Supplementary File S2 (evidence bundle ZIP). The raw prompt pool, full T5-PAWS attack-generation trace, and LLaMA-3-8B full paired arrays are not included in the current bundle because complete source-level licensing and trace information are unavailable. The LLaMA-3-8B paired row in Table 6 is therefore aggregate diagnostic output and is not used as independently deposited paired evidence. Operational deployment keys are not provided; the released materials use research-only seeds or non-deployment identifiers for reproducibility. Further inquiries about the raw attack-generation trace can be directed to the corresponding author, subject to licensing and storage constraints.

**Conflicts of Interest**

The authors declare no conflicts of interest.

**Abbreviations**

The following abbreviations are used in this manuscript: AEG, Adaptive Entropy Gate and adaptive entropy-gated watermarking; BPE, byte-pair encoding; CSPRNG, cryptographically secure pseudorandom number generator; FPR, false-positive rate; KGW, Kirchenbauer–Geiping–Wen watermarking baseline; LLM, large language model; NF4, 4-bit NormalFloat quantization; OM, oracle-minimum attack ensemble; PPL, perplexity; PRF, pseudorandom function; ROC, receiver operating characteristic; SBERT, Sentence-BERT; SIR, Zhao et al. context-dependent watermarking baseline; TPR, true-positive rate.

**REFERENCES**

Aaronson, S. (2022, December). Watermarking GPT outputs. Shtetl-Optimized Blog. https://scottaaronson.blog/?p=6865. Accessed June 10, 2026.

BigScience Workshop. (2022). BLOOM: A 176B-parameter open-access multilingual language model. arXiv. https://arxiv.org/abs/2211.05100. Accessed June 10, 2026.

Chang, Y., Krishna, K., Houmansadr, A., Wieting, J. F., & Iyyer, M. (2024). PostMark: A robust blackbox watermark for large language models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 8969-8987). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.506

Christ, M., Gunn, S., & Zamir, O. (2023). Undetectable watermarks for language models. arXiv. https://arxiv.org/abs/2306.09194. Accessed June 10, 2026.

Dabiriaghdam, A., & Wang, L. (2025). SimMark: A robust sentence-level similarity-based watermarking algorithm for large language models. arXiv. https://arxiv.org/abs/2502.02787. Accessed June 10, 2026.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. arXiv. https://arxiv.org/abs/2305.14314. Accessed June 10, 2026.

Hou, J., Zhang, J., Ma, Z., & Xu, X. (2023). MPAC: A multi-party watermarking scheme for AI-generated text. arXiv. https://arxiv.org/abs/2307.01318. Accessed June 10, 2026.

Hu, X., Chen, P.-Y., & Ho, T.-Y. (2023). RADAR: Robust AI-text detection via adversarial learning. In Advances in Neural Information Processing Systems (Vol. 36). https://proceedings.neurips.cc/paper_files/paper/2023/hash/30e15e5941ae0cdab7ef58cc8d59a4ca-Abstract-Conference.html. Accessed June 10, 2026.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Renard Lavaud, L., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). Mistral 7B. arXiv. https://arxiv.org/abs/2310.06825. Accessed June 10, 2026.

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 (Vol. 202, pp. 17061-17084). PMLR. https://proceedings.mlr.press/v202/kirchenbauer23a.html. Accessed June 10, 2026.

Kirchenbauer, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., Fernando, K., Saha, A., Goldblum, M., & Goldstein, T. (2024). On the reliability of watermarks for large language models. International Conference on Learning Representations. https://proceedings.iclr.cc/paper_files/paper/2024/hash/d78e9e4316e1714fbb0f20be66f8044c-Abstract-Conference.html. Accessed June 10, 2026.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In Advances in Neural Information Processing Systems (Vol. 36). https://openreview.net/forum?id=WbFhFvjjKj. Accessed June 10, 2026.

Kuditipudi, R., Thickstun, J., Hashimoto, T., & Liang, P. (2023). Robust distortion-free watermarks for language models. arXiv. https://arxiv.org/abs/2307.15593. Accessed June 10, 2026.

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). Pointer sentinel mixture models. arXiv. https://arxiv.org/abs/1609.07843. Accessed June 10, 2026.

Meta AI. (2024). The Llama 3 herd of models. arXiv. https://arxiv.org/abs/2407.21783. Accessed June 10, 2026.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. In Proceedings of the 40th International Conference on Machine Learning (pp. 24950-24962). PMLR. https://proceedings.mlr.press/v202/mitchell23a.html. Accessed June 10, 2026.

Qwen Team. (2024). Qwen2 technical report. arXiv. https://arxiv.org/abs/2407.10671. Accessed June 10, 2026.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1-67. http://jmlr.org/papers/v21/20-074.html. Accessed June 10, 2026.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (pp. 3982-3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text really be detected? arXiv. https://arxiv.org/abs/2303.11156. Accessed June 10, 2026.

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT: Building open translation services for the world. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (pp. 479-480). https://aclanthology.org/2020.eamt-1.61/. Accessed June 10, 2026.

Verma, A., Phan, H., & Trivedi, S. (2026). Watermarking degrades alignment in language models: Analysis and mitigation. Transactions on Machine Learning Research. https://openreview.net/forum?id=w2ATKQcfWx. Accessed June 10, 2026.

Yoo, K., Ahn, W., Jang, J., & Kwak, N. (2023). Robust multi-bit natural language watermarking through invariant features. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (pp. 2092-2115). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.117

Zhang, Y., Baldridge, J., & He, L. (2019). PAWS: Paraphrase adversaries from word scrambling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1298-1308). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1131

Zhao, L., Wang, W., Kuang, K., Li, H., Wu, F., & Xiao, J. (2023). Provable robust watermarking for AI-generated text. arXiv. https://arxiv.org/abs/2306.17439. Accessed June 10, 2026.