When Do LLM Watermarks Lose Information? A Diagnostic Study under Controlled Rewriting

Anonymous ACL submission

Abstract

Text watermarking provides provenance signals for LLM-generated content, yet reliability after rewriting is usually evaluated with AUROC or TPR. These scores do not show when a deployed detector's statistic has lost most of its watermark information. We introduce CAP-Diag, a diagnostic protocol for measuring detector-statistic- accessible mutual information under controlled rewrites. Across two English corpora and three watermark families, CAP-Diag reveals a regime- and length-coupled bottleneck: compressed fact-skeleton rewrites retain substantially less watermark information than ordinary paraphrases. Based on this observation, we propose the Low-Capacity Rewrite Attack, which rewrites watermarked long-form text into the diagnosed low-information regime and reduces accessible information by 77–91%. We further derive CAP-Triage, an abstention readout for KGW and UW detectors when the chosen statistic is likely uninformative. Additional reference contrasts show that the bottleneck is mechanism-dependent rather than universal. Code is available at https://anonymous.4open.science/r/CAP-Diag-3E5E.

1 Introduction

Text watermarking for large language models aims to identify machine-generated content by embedding provenance signals during generation (Liu et al., 2024b; Kirchenbauer et al., 2023; Zhao et al., 2024; Kuditipudi et al., 2024). Representative token-level methods include KGW, Unigram-Watermark (UW), and EXP/Gumbel watermarks (Kirchenbauer et al., 2023; Zhao et al., 2024; Kuditipudi et al., 2024). In deployment, however, detectors often observe rewritten text rather than the original generation: a response may be paraphrased, compressed, summarized, or elaborated by a downstream user. Existing evaluations therefore report detection metrics such as AUROC and TPR at fixed FPR, together with robustness under rewriting and removal attacks (Tu et al., 2024; Pan et al., 2024; Krishna et al., 2023; Sadasivan et al., 2025; Chang et al., 2024a; Wu and Chandrasekaran, 2024).

These metrics leave a complementary question unanswered: after a utility-preserving rewrite, how much watermark information remains in the statistic that a deployed detector actually consumes? A detector may score well in one regime while its statistic becomes nearly uninformative in another. In that case, retuning the threshold cannot recover reliable decisions for any classifier that only uses the same statistic.

We introduce CAP-Diag, a diagnostic protocol for measuring detector-statistic-accessible watermark information under controlled rewriting. For a fixed detector statistic S, CAP-Diag estimates the mutual information between the watermark flag and S across a ladder of utility-preserving rewrite regimes, and interprets low information through a chosen-statistic error floor. The diagnostic is statistic-conditional: it measures what is available to the specified statistic, not what might be recovered by richer detectors.

Across KGW, UW, and EXP/Gumbel watermarks on synthetic and Wikipedia naturalistic English, CAP-Diag identifies a regime- and length-coupled low-information bottleneck. Compressed fact-skeleton rewrites retain far less detector-accessible information than ordinary paraphrases, with R1/R0 mutual-information ratios of roughly 4–15$\times$ across the three token-level families. Length is a central confound, but length-matched controls show that token count alone does not explain the gap. This complements long-text reliability analyses (Kirchenbauer et al., 2024): long observed spans can remain detectable, while compressed rewrites can leave the chosen statistic with too little accessible evidence.

This bottleneck leads to a direct attack conse-

quence. The Low-Capacity Rewrite Attack rewrites otherwise detectable watermarked long-form text into the diagnosed low-information regime while preserving the source facts and passing the utility gate. Under the corresponding low-capacity reference readout, it reduces detector-accessible mutual information by 77–91% across KGW, UW, and EXP/Gumbel. Unlike generic rewrite baselines, LowCap is diagnosis-guided: it targets the regime that CAP-Diag identifies as low-information for the chosen statistic.

The same diagnosis also supports a deployment readout. For KGW and UW greenlist detectors, CAP-Triage abstains when an observable proxy indicates that a document is dominated by low-information segments. Reference contrasts then delimit the scope of the result: a sentence-level semantic reference (Hou et al., 2024) exhibits the same R0 versus R1 pattern, whereas a content-presence reference (Chang et al., 2024b) does not. The bottleneck is therefore mechanism-dependent rather than universal.

Figure 1 summarizes this workflow: controlled rewriting and utility gating produce accepted rewrite regimes, CAP-Diag identifies the R0 low-information bottleneck, and the diagnosis supports CAP-Triage and LowCap.

Our contributions are summarized as follows:

- Diagnosis: We introduce CAP-Diag, a statistic-conditional protocol for measuring detector-accessible watermark information, and reveal an R0 low-information bottleneck under controlled rewrite.
- Attack: Based on this bottleneck, we propose the Low-Capacity Rewrite Attack, a diagnosis-guided reachability attack that rewrites watermarked long-form text into the diagnosed low-information regime.
- Readout and scope: We derive CAP-Triage as an abstention readout for low-information cases and use reference contrasts to show that the bottleneck is mechanism-dependent rather than universal.

2 Background

2.1 Watermarking and Detector Statistics

A text watermark consists of a generation-time embedding rule and a detection-time statistic. CAP-Diag treats the scalar statistic S as the detector's observable watermark evidence.

KGW (Kirchenbauer et al., 2023) and Unigram-Watermark (UW) (Zhao et al., 2024) are greenlist watermarks. KGW derives a secret green list from the recent context, whereas UW uses a fixed secret green list across positions. In both families, the detector counts green-token occurrences and normalizes the count as

$$z = \frac{n_{\text{green}} - \gamma T}{\sqrt{T\gamma(1-\gamma)}},$$

where T is the observed length and γ is the greenlist fraction. We refer to these as greenlist z-statistic detectors.

EXP/Gumbel watermarks (Kuditipudi et al., 2024) use a different token-level statistic. Instead of counting green-list membership, the detector aggregates a secret-keyed per-token score,

$$S_T = \sum_{t=1}^{T} -\log(1 - r_{x_t}),$$

where r_{x_t} is the key-derived value associated with the observed token. Thus, EXP/Gumbel is a token-level distribution-deviation statistic but not a greenlist z-statistic detector. Our diagnostic and attack analyses cover all three token-level families; CAP-Triage is scoped to the two greenlist z-statistic detectors.

2.2 Controlled Rewrite Regimes

We study watermark information after utility-preserving rewrites. A single paraphrase setting is too coarse for this purpose, because short fact-preserving summaries, ordinary paraphrases, looser rewrites, and long elaborations can expose different amounts of detector evidence. We therefore use a controlled rewrite ladder.

R0 is a compressed fact-skeleton rewrite that preserves the source's key facts in a short form. R1 is an ordinary short paraphrase. R1.5 permits looser surface variation while remaining utility-preserving. R2 is a longer elaboration that retains the source content and gives the detector a larger observed span. These regimes are controlled operating points, not a taxonomy of all possible rewrites.

A rewrite is retained only if it passes a utility-preserving gate $\mathcal{G}$. The gate checks bidirectional non-contradiction between the source and rewrite, preservation of the source's fact skeleton, and membership in the target length band for the regime. Prompt templates, gate implementation details, and gate-policy sensitivity checks are reported in the appendices.

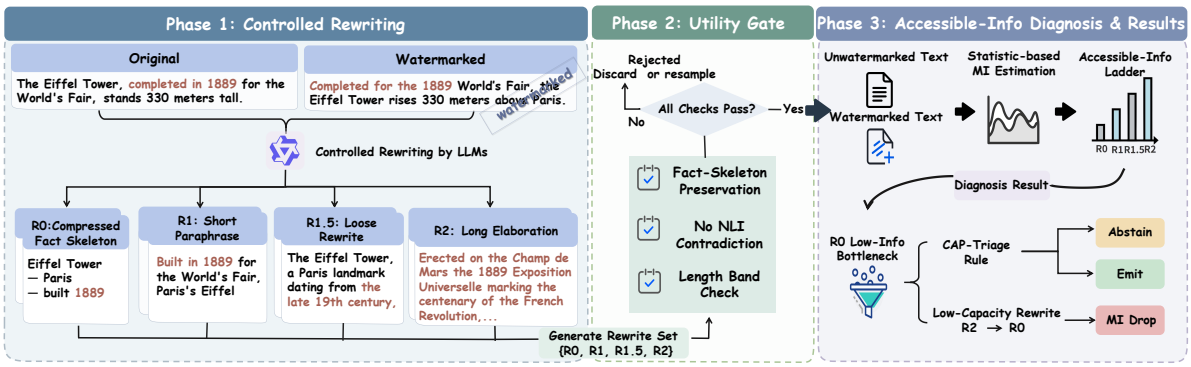

Figure 1: Overview of CAP-Diag and its downstream consequences. Controlled rewriting generates candidate regimes R0, R1, R1.5, and R2. The utility gate retains rewrites that preserve the fact skeleton, avoid NLI contradiction, and satisfy the target length band. CAP-Diag estimates detector-statistic-accessible information from matched watermarked and unwatermarked rewrites, producing an accessible-information ladder and identifying the R0 low-information bottleneck. The diagnosis then supports CAP-Triage abstention and the Low-Capacity Rewrite Attack.

3 Detector-Statistic-Accessible Information

This section defines CAP-Diag and applies it to the controlled rewrite ladder. The main question is whether compressed fact-skeleton rewrites leave the chosen detector statistic with less usable watermark information than ordinary paraphrases.

3.1 Definition

Let $W \in \{0, 1\}$ denote whether a text is watermarked, and let S denote the scalar statistic consumed by a detector after rewriting. CAP-Diag estimates, within each rewrite regime, how much information S retains about W. This quantity is statistic-conditional: it measures information available to the specified detector statistic, not information recoverable by an arbitrary full-text detector.

Given balanced samples of watermarked and unwatermarked texts, we estimate

$$\begin{aligned}\hat{I}(W;S) = \hat{H}_{\mathrm{MM}}(\mathrm{pooled}) \\ - \pi_{\mathrm{wm}}\hat{H}_{\mathrm{MM}}(S_{\mathrm{wm}}) \\ - (1-\pi_{\mathrm{wm}})\hat{H}_{\mathrm{MM}}(S_{\mathrm{no}}),\end{aligned}$$

where $\hat{H}_{\mathrm{MM}}$ is a Miller–Madow-corrected histogram entropy estimator and π_{wm} is the watermarked fraction. The main experiments use $\pi_{\mathrm{wm}} = 0.5$ and source-paired bootstrap intervals.

We use $\hat{I}(W;S)$ as a separability diagnostic for the chosen statistic. It is not a Shannon channel-capacity estimate and does not upper-bound richer detectors. Our conclusions rely on the stable R0 versus higher-regime separation, paired uncertainty intervals, and standard detection metrics reported on the same cells. Estimator details, prior-shift variants, classifier lower bounds, and the full algorithm are given in Appendices A, A.1, and A.2.

Proposition 1 (Chosen-statistic error floor)**.** Let $W \in \{0, 1\}$ have prior π, and let $\hat{W} = h(S)$ be any classifier that consumes only the chosen statistic S. If $P_e = \Pr[\hat{W} \neq W]$, then

$$P_e \;\geq\; H_b^{-1}\big(H(W) - I(W;S)\big),$$

where H_b^{-1} is the inverse binary entropy on $[0, 1/2]$.

Prop. 1 gives the operational meaning of small detector-statistic-accessible MI: when $I(W;S)$ is small, no threshold or classifier using only S can recover reliable detection. The full derivation is provided in Appendix A.

3.2 Experimental Setup

We evaluate CAP-Diag on two English corpora: a synthetic corpus with 200 fact-skeleton episodes and a naturalistic corpus with 300 Wikipedia passages from fifteen domains. R0 and R1 are generated for the full naturalistic corpus; R1.5 and R2 are evaluated on a 150-source subset. Corpus construction details are reported in Appendix B.

The primary generator is Qwen2.5-7B-Instruct (Hui et al., 2024); a small cross-generator validation with Llama-3.1-8B-Instruct (Grattafiori et al., 2024) is reported in Appendix B.1. We evaluate KGW, UW, and EXP/Gumbel as the three token-level watermark families. KGW and UW are evaluated across a strength sweep, while EXP/Gumbel is evaluated at its strongest

Table 1: Independent LLM audit on accepted rewrites. DeepSeek v4-pro is separate from the Qwen2.5-7B-Instruct generator used for rewrite generation.

Class	n	Passed	Pass rate
R0	100	99	99%
R1	100	100	100%
R1.5	100	100	100%
R2	100	100	100%
LowCap attack	100	99	99%

setting. All reported cells match watermarked and unwatermarked samples under the same rewrite regime and utility gate, and report MI, AUROC, and TPR at fixed FPR on the same cells. Implementation details, prompts, source identifiers, aggregate detector scores, and key-sensitivity checks are provided in the appendices.

Rewrite-quality audit. Because CAP-Diag compares detector statistics across accepted rewrites, we audit whether those rewrites preserve source utility. The audit uses DeepSeek v4-pro, independent from the Qwen2.5-7B-Instruct generator, and checks fact preservation, factual drift, and main meaning. As shown in Table 1, accepted rewrites pass this audit in 498 of 500 cases across the four rewrite regimes and the Low-Capacity attack condition.

3.3 Results

Figure 2 and Table 2 report the main accessible-information ladder. Across all three token-level families, R0 is consistently the low-information regime: ordinary paraphrases and long elaborations expose substantially more detector-statistic-accessible information than compressed fact-skeleton rewrites. The naturalistic R1.5/R2 cells are evaluated on the 150-source subset and serve as supporting evidence for the ladder.

Because R1.5 has a more selective utility gate, we treat it as supporting evidence and use the R0 versus R1 contrast as the primary diagnostic comparison. AUROC and TPR on the same cells show the same R0 failure pattern and are reported in the appendix. Estimator, detector-aggregation, gate-policy, and key-sensitivity checks preserve the R0 $\ll$ R1 ordering; Appendix D reports the full sensitivity results.

3.4 Length Control

R0 rewrites are shorter than R1 rewrites, making length the central observable confound. We analyze this confound by comparing R0 and R1 within overlapping length buckets and by generating an R1-short variant constrained to the R0 length band. The goal is not a full causal decomposition of every surface attribute, but to test whether the diagnosed R0 bottleneck is only a token-count artifact.

Within the overlapping length window, R1 remains several times more informative than R0. Length is therefore tightly coupled with the diagnosed bottleneck and provides a useful observable proxy for the abstention readout in Sec. 5.1. Detailed bucket-level statistics and source-paired tests are reported in Appendix D.

Figure 3 summarizes the length-band control: R1-short occupies the same length band as R0, yet retains more detector-accessible information than R0. At controlled length, the source-paired R1-short/R0 MI ratio remains above one for all three token-level families: $1.57\times$ for KGW, $2.71\times$ for UW, and $3.20\times$ for EXP/Gumbel, with confidence intervals excluding one in each case. These ratios are smaller than the unmatched R1-original/R0 ratios, but they show that constraining R1 to the R0 length band does not eliminate the regime gap.

R0 therefore does not fail merely because it is short. Short length and compressed fact-skeleton rewriting jointly reduce the evidence available to the detector statistic. CAP-Diag identifies a length-coupled low-information regime, not a full causal decomposition of every surface factor inside that regime. This complements long-text reliability analyses (Kirchenbauer et al., 2024): long observed watermarked spans can remain detectable, while compressed rewrites can leave the chosen statistic with too little accessible information. This diagnosed low-information regime is the basis for the attack consequence studied next.

4 Low-Capacity Rewrite Attack

Section 3 identifies R0 as a low-information evidence regime for the studied token-level detector statistics. The Low-Capacity Rewrite Attack tests whether this regime is reachable from otherwise detectable watermarked text. Given a watermarked long-form R2 rewrite, the attack rewrites it into an R0-style compressed fact skeleton while preserving the source facts. It is a diagnosis-guided reachability test, not a claim to be the strongest possible removal attack.

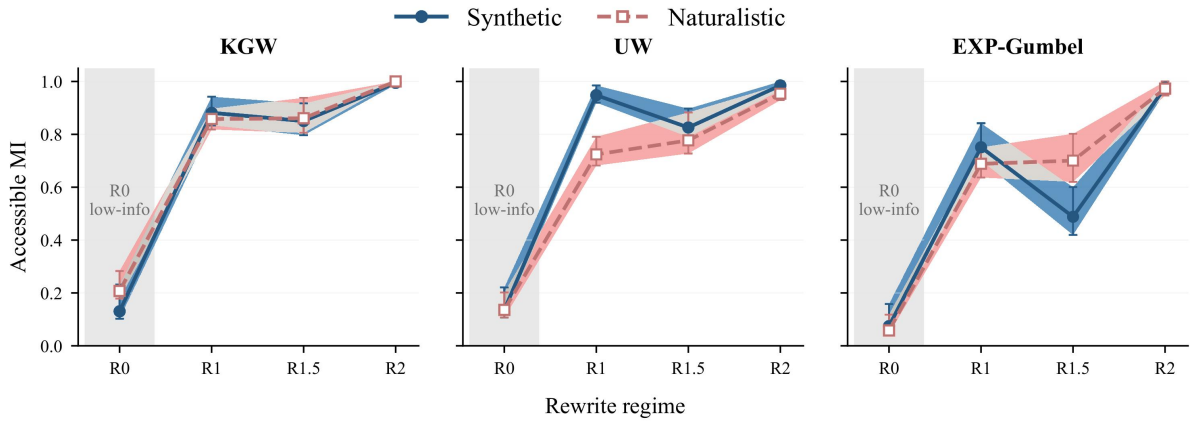

Figure 2: Accessible-information ladder. CAP-Diag estimates detector-statistic-accessible MI for KGW, UW, and EXP/Gumbel across R0, R1, R1.5, and R2. Lines show mean MI with 95% source-paired bootstrap intervals; the gray band marks the diagnosed low-information R0 regime. Naturalistic R1.5/R2 cells use the 150-source subset.

Table 2: Detector-statistic-accessible MI by rewrite regime. The table reports MI, 95% source-paired bootstrap intervals, and mean token counts. Bold values mark the diagnosed low-information R0 regime; naturalistic R1.5/R2 cells use the 150-source subset.

Corpus	Regime	KGW			UW			EXP-Gumbel		
		MI	95% CI	Tok.	MI	95% CI	Tok.	MI	95% CI	Tok.
Synthetic										
	R0	**0.131**	[0.103, 0.232]	33	**0.136**	[0.107, 0.221]	32	**0.074**	[0.062, 0.158]	32
	R1	0.882	[0.834, 0.942]	52	0.948	[0.920, 0.985]	50	0.751	[0.701, 0.842]	50
	R1.5	0.850	[0.796, 0.917]	48	0.826	[0.789, 0.897]	48	0.488	[0.419, 0.600]	44
	R2	0.995	[0.988, 1.000]	101	0.985	[0.965, 1.000]	113	0.977	[0.960, 1.000]	142
Naturalistic										
	R0	**0.207**	[0.179, 0.283]	60	**0.136**	[0.107, 0.202]	61	**0.058**	[0.046, 0.118]	61
	R1	0.858	[0.818, 0.897]	87	0.724	[0.682, 0.791]	86	0.689	[0.637, 0.751]	89
	R1.5	0.861	[0.806, 0.938]	104	0.777	[0.728, 0.883]	95	0.701	[0.620, 0.802]	98
	R2	1.000	[1.000, 1.000]	173	0.954	[0.929, 0.987]	155	0.973	[0.947, 1.000]	179

4.1 Setup

The attacker is given a watermarked R2 rewrite of a source document. The attack applies an R0-style rewrite instruction, keeps the result only if it passes the same utility gate used in CAP-Diag, and leaves the detector unchanged. No additional watermark is applied.

We evaluate the attack on 100 source episodes for each of KGW, UW, and EXP/Gumbel. Each family uses the same detector statistic, MI estimator, and source-paired uncertainty procedure as the main diagnostic analysis.

The main readout uses an R0-reference null: post-rewrite watermarked outputs are compared against unwatermarked R0 outputs that match the low-capacity length and surface regime. We also compare LowCap R0 with generic rewrite baselines, including ordinary R1 paraphrase and length-constrained R1-short, under the same detector-statistic readout.

4.2 Results

Figure 4 shows that no-attack R2 is nearly separable, while all rewrite conditions reduce separability under the R0-reference readout. Generic R1 baselines can erase token-level evidence more aggressively in some cells, but they measure broad rewrite fragility. LowCap R0 instead tests whether the diagnosed low-information R0 regime is reachable from detectable watermarked text.

Table 3 reports the same readout with standard metrics and MI drop. LowCap R0 reduces detector-statistic-accessible information by 81.4% $[69.6, 84.6]$ for KGW, 77.1% $[66.2, 82.6]$ for UW, and 90.7% $[77.2, 93.0]$ for EXP/Gumbel. Post-attack MI falls to 0.134, 0.164, and 0.093, respectively; EXP/Gumbel AUROC falls from 1.00 before the attack to 0.64 under the R0-reference read-

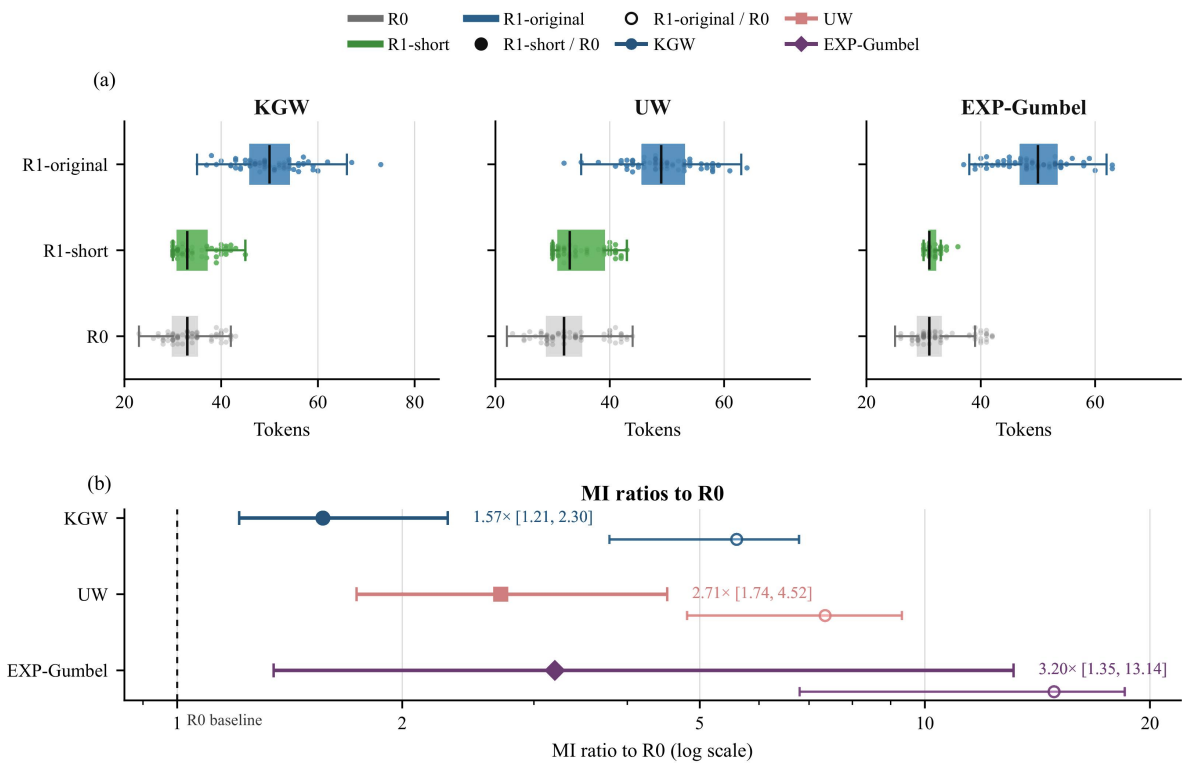

Figure 3: Length-control analysis. Panel (a) compares token-count distributions for R0, R1-short, and R1-original. Panel (b) reports MI ratios to R0 on a log scale; filled markers show length-controlled R1-short/R0 contrasts and open markers show unmatched R1-original/R0 contrasts. The dashed line marks parity with R0.

Table 3: Low-Capacity Rewrite Attack under the R0-reference readout. Values are computed on the same cells as Fig. 4. MI drop is relative to the no-attack R2 condition.

Rewrite	KGW			UW			EXP-Gumbel		
	AUROC	TPR@5	MI drop	AUROC	TPR@5	MI drop	AUROC	TPR@5	MI drop
Detector signal after rewrite									
No attack R2	1.00	1.00	0%	1.00	1.00	0%	1.00	1.00	0%
R1 paraphrase	0.51	0.11	98%	0.50	0.14	94%	0.53	0.08	98%
R1-short	0.55	0.12	99%	0.43	0.08	93%	0.61	0.12	94%
LowCap R0	0.76	0.38	81%	0.76	0.43	77%	0.64	0.16	91%

out.

The attacked rewrites remain within the same utility constraints used by CAP-Diag: the source facts are preserved and the rewrites pass the bidirectional NLI gate. Thus, the low-information regime is not only an evaluation cell; it is reachable from otherwise detectable watermarked text.

5 Deployment Readout and Mechanism Boundary

The Low-Capacity Rewrite Attack exposes a deployment problem: a detector may be asked to decide on text whose observed statistic is uninformative. We study an abstention readout for this setting and then test the mechanism boundary of the bottleneck.

5.1 CAP-Triage as Abstention

Setup and metric. CAP-Triage is an abstention readout rather than a recovery detector: it withholds a binary decision when a source-free proxy indicates an R0-like low-information segment. We instantiate it for KGW and UW greenlist z-statistic detectors using document length as the proxy. Coverage is the fraction of watermarked documents on which the system emits a decision; false-confident rate is the frequency of high-confidence emissions on unwatermarked low-information documents. Oracle-regime and source-relative perplexity proxies are appendix reference readouts.

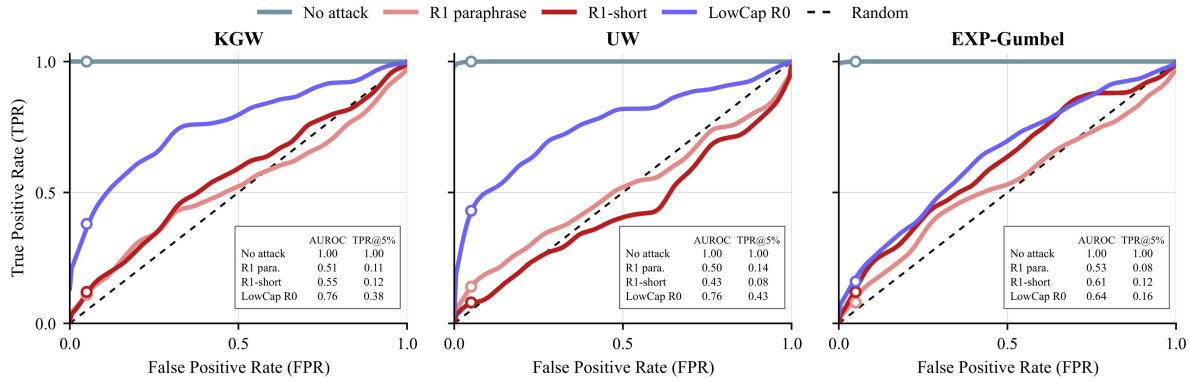

Figure 4: Low-Capacity Rewrite Attack and generic rewrite baselines. Each panel reports ROC curves for one detector family under the same R0-reference readout. LowCap R0 is the diagnosis-induced condition; R1 paraphrase and R1-short are generic rewrite baselines. Insets report AUROC and TPR at 5% FPR.

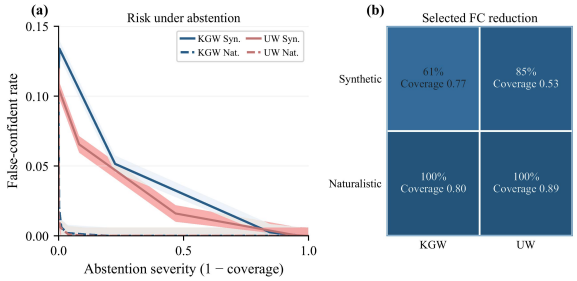

Figure 5: CAP-Triage as an abstention readout. Panel (a) plots false-confident rate against abstention severity ($1 - \text{coverage}$) for KGW and UW on held-out splits. Panel (b) summarizes the pre-specified operating point: the largest false-confident-emission reduction subject to at least 50% retained coverage. Each cell reports the selected reduction and coverage.

We summarize each frontier with a pre-specified rule: among thresholds on the fixed abstention grid, choose the largest false-confident-rate reduction subject to at least 50% coverage. This summarizes the frontier rather than defining a universal deployment threshold. We evaluate the rule on held-out synthetic sources and held-out Wikipedia domains. In Fig. 5, the x-axis is abstention severity, $1 - \text{coverage}$.

Risk–coverage result. Figure 5 shows that stronger abstention lowers false-confident emissions for both KGW and UW. Under the pre-specified summary rule, CAP-Triage reduces false-confident emissions by 61% at 77% coverage for KGW and by 85% at 53% coverage for UW on the synthetic held-out split, with the naturalistic held-out split summarized in the selected-gain matrix. Thus the readout reduces false-confident emissions by abstaining on low-information cases.

5.2 Mechanism Boundary

These results are not a universal impossibility claim. CAP-Diag is conditional on the chosen statistic, and CAP-Triage also requires a useful proxy. We use reference contrasts to test whether the R0/R1 pattern appears under other evidence mechanisms.

Mechanism contrasts. Table 4 shows that a sentence-level semantic reference (Hou et al., 2024) also has an R0/R1 gap, with R1/R0 MI ratios of roughly $9.7\times$ on the synthetic subset and $6.8\times$ on the naturalistic subset. Thus the pattern is not limited to token-level greenlist counting. By contrast, a content-presence reference (Chang et al., 2024b) stays near $1.0\times$ and remains high-AUROC in both regimes. The bottleneck therefore appears for evidence sensitive to rewrite-induced distributional or semantic changes, but not for content-survival evidence. These references are mechanism contrasts, not full reproductions of the original systems.

EXP/Gumbel scope. EXP/Gumbel separates diagnosis from deployment readout. LowCap reduces EXP/Gumbel MI in Sec. 4.2, but the current length proxy does not support CAP-Triage: the held-out sample has no false-confident budget on unwatermarked R0 documents, and attacked documents are not shorter than canonical R1 documents. We therefore scope CAP-Triage to KGW and UW, while treating EXP/Gumbel as evidence that other statistics may require different proxies.

6 Related Work

6.1 LLM Watermarking

Text watermarking for LLMs embeds provenance signals during generation and detects them with a watermark statistic. Existing surveys organize this

Table 4: Mechanism-boundary contrasts. Ratios compare accessible MI in R1 against R0; values near one indicate no R0/R1 bottleneck for that reference statistic. N/A marks statistics used only for mechanism-boundary analysis, not CAP-Triage.

Statistic	Evidence type	Syn. R1/R0	Nat. R1/R0	CAP-Triage
KGW	greenlist z	$6.7\times$	$4.1\times$	yes
UW	greenlist z	$7.0\times$	$5.3\times$	yes
EXP-Gumbel	token score sum	$10.1\times$	$11.9\times$	no
Sentence-level semantic ref.	semantic hash	$9.7\times$	$6.8\times$	N/A
Content-presence ref.	content survival	$\approx 1.0\times$	$\approx 1.0\times$	N/A

area around detectability, quality, payload, robustness, security, and deployment assumptions (Liu et al., 2024b). The three token-level families evaluated in this paper are defined in Sec. 2. Related work studies undetectable watermarking, unbiased or debiased designs, scalable sampling-time watermarking, statistical coupling-based constructions, and designs that revisit the detectability–robustness–quality trade-off (Christ et al., 2024; Hu et al., 2024; Xie et al., 2025; Dathathri et al., 2024; Giboulot and Furon, 2024). CAP-Diag does not propose a new embedding rule; it measures how much information a chosen detector statistic retains after controlled utility-preserving rewrite.

Evaluation frameworks and toolkits compare watermarking methods with quality, detection, and robustness metrics (Tu et al., 2024; Pan et al., 2024). Theoretical and statistical work studies watermark guarantees under quality, robustness, edit, and attacker constraints (Pang et al., 2024; Zhang et al., 2024; Li et al., 2025). Recent security work also shows that watermark deployment can be probed in black-box settings, making exposed statistics and query access part of the practical threat surface (Gloaguen et al., 2025). These lines motivate our statistic-conditional view: CAP-Diag measures the information available to a specified statistic, rather than characterizing all possible detectors. Semantic and structure-aware watermarks provide different evidence mechanisms (Liu et al., 2024a; Ren et al., 2024); Sec. 5.2 uses reference contrasts to delimit the mechanism scope of our findings.

6.2 Watermark Robustness

A useful watermark detector should remain reliable after utility-preserving transformations. Prior work studies paraphrase and recursive rewriting, stress tests for detector reliability, smoothing attacks, color-aware substitutions, watermark stealing, and targeted self-information rewrite attacks (Krishna et al., 2023; Sadasivan et al., 2025; Chang et al., 2024a; Wu and Chandrasekaran, 2024; Jovanović et al., 2024; Cheng et al., 2025). These works ask whether a transformation or an adversary can weaken, remove, spoof, or bypass watermark evidence.

Recent robustness benchmarks systematize multiple watermarking methods and removal attacks, showing that robustness depends strongly on design choices and evaluation protocols (Liang et al., 2025). Low-Capacity rewriting asks a more specific question: once CAP-Diag identifies a low-information rewrite regime, can that regime be reached from otherwise detectable watermarked text? It is therefore a diagnosis-guided reachability attack, not a claim to be the strongest possible removal attack. Cross-lingual watermark robustness is closest in structure: it diagnoses a transformation-induced failure mode, derives an attack, and studies how watermark designs might respond (He et al., 2024). Our work follows this diagnostic-to-consequence pattern, but measures detector-statistic-accessible information under controlled rewrite and connects the diagnosed bottleneck to LowCap and CAP-Triage.

7 Conclusion

We introduced CAP-Diag, a statistic-conditional diagnostic for measuring watermark information available to a chosen detector statistic under controlled rewrite. Across three token-level watermark families, CAP-Diag reveals a regime- and length-coupled low-information bottleneck in compressed fact-skeleton rewrites. This bottleneck induces the Low-Capacity Rewrite Attack, which reduces accessible information by 77–91% under the low-capacity reference readout, and CAP-Triage turns the diagnosis into an abstention readout for KGW and UW. Reference contrasts further show that the bottleneck is mechanism-dependent rather than universal.

Limitations

CAP-Diag is statistic-conditional. It measures the information retained by a chosen detector statistic S, not the information that could be recovered by an arbitrary learned detector or by access to the full text. The Fano-style error floor therefore applies to classifiers that consume the specified statistic. A richer statistic may expose more watermark signal, and our adaptive-detector checks should be read as sensitivity analyses rather than as an exhaustive search over all possible detectors. The estimates are also prior-conditional: the main experiments use a balanced watermarked/unwatermarked evaluation prior, while deployments with different base rates would require the corresponding prior adjustment.

The empirical scope is English and rewrite-centered. We evaluate three token-level watermark families on synthetic fact-skeleton episodes and Wikipedia passages, with Qwen2.5-7B-Instruct as the primary generator and a smaller cross-generator validation in the appendix. We do not evaluate multilingual settings, conversational data, code, or domain-specific corpora beyond Wikipedia. The naturalistic corpus is also uneven across domains, so the per-domain results are best read as a robustness check rather than as a complete domain-generalization study.

The mechanism-boundary experiments are deliberately bounded. We use one sentence-level semantic reference and one content-presence reference to test whether the R0 versus R1 bottleneck appears under different detector statistics. These references are mechanism contrasts rather than full reproductions of the original systems. The results therefore support a narrower conclusion: the bottleneck is evident for the token-level distribution-deviation statistics studied here, appears in one sentence-level semantic reference, and does not appear in one content-presence reference. They do not establish a mechanism-class theorem.

CAP-Triage is also narrower than CAP-Diag. The diagnostic and attack analyses cover KGW, UW, and EXP/Gumbel, but the abstention readout is evaluated only for KGW and UW greenlist detectors. EXP/Gumbel exhibits the Low-Capacity Rewrite Attack, but the current length-based proxy does not provide a useful abstention signal for that family. CAP-Triage should therefore be interpreted as a deployment readout for a specific detector/proxy pairing, not as a general defense or a method for recovering watermark information once the chosen statistic has become uninformative.

Ethical Considerations

This work studies a failure mode of watermark detection under meaning-preserving rewrite. The Low-Capacity Rewrite Attack does not require privileged access to the detector or to the watermark secret; it formalizes a risk that is already present whenever users can compress or paraphrase watermarked text. Reporting the attack may make this risk more explicit, but it also clarifies when watermark detectors should not be trusted to return a definitive provenance decision.

The intended use of CAP-Diag and CAP-Triage is conservative deployment. CAP-Diag identifies regimes in which the chosen detector statistic contains too little accessible information for reliable decisions, and CAP-Triage turns this diagnosis into abstention for the KGW and UW settings studied here. Such abstention can reduce overconfident provenance claims, but it does not replace broader safeguards for content attribution, provenance tracking, or user transparency.

Watermark detectors can create harms when overtrusted. False positives may attribute human-written text to a machine system, while false negatives may miss watermarked content after rewriting. These risks may vary across languages, domains, writing styles, and demographic or social contexts not covered by our English Wikipedia-centered evaluation. We therefore treat the results as evidence about specific detector statistics under controlled rewrite, not as a guarantee of uniform performance in deployment.

References

Hongyan Chang, Hamed Hassani, and Reza Shokri. 2024a. Watermark smoothing attacks against language models. *arXiv preprint arXiv:2407.14206*.

Yapei Chang, Kalpesh Krishna, Amir Houmansadr, John Frederick Wieting, and Mohit Iyyer. 2024b. Postmark: A robust blackbox watermark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8969–8987.

Yixin Cheng, Hongcheng Guo, Yangming Li, and Leonid Sigal. 2025. Revealing weaknesses in text watermarking through self-information rewrite attacks. In *Forty-second International Conference on Machine Learning*.

Miranda Christ, Sam Gunn, and Or Zamir. 2024. Undetectable watermarks for language models. In *Proceedings of the 37th Conference on Learning Theory*.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, and 1 others. 2024. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823.

Eva Giboulot and Teddy Furon. 2024. Watermax: breaking the llm watermark detectability-robustness-quality trade-off. *Advances in Neural Information Processing Systems*, 37:18848–18881.

Thibaud Gloaguen, Nikola Jovanović, Robin Staab, and Martin Vechev. 2025. Black-box detection of language model watermarks. In *International Conference on Learning Representations*, volume 2025, pages 77539–77562.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zhiwei He, Binglin Zhou, Hongkun Hao, Aiwei Liu, Xing Wang, Zhaopeng Tu, Zhuosheng Zhang, and Rui Wang. 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 (Volume 1: Long Papers)*, pages 4115–4129.

Abe Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. 2024. Semstamp: A semantic watermark with paraphrastic robustness for text generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4067–4082.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. 2024. Unbiased watermark for large language models. In *International Conference on Learning Representations*, volume 2024, pages 45408–45436.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Nikola Jovanović, Robin Staab, and Martin Vechev. 2024. Watermark stealing in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. 2024. On the reliability of watermarks for large language models. In *International Conference on Learning Representations*, volume 2024, pages 49660–49704.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in neural information processing systems*, 36:27469–27500.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2024. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research*.

Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. 2025. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *The Annals of Statistics*, 53(1):322–351.

Jiacheng Liang, Zian Wang, Spencer Hong, Shouling Ji, and Ting Wang. 2025. Watermark under fire: A robustness evaluation of LLM watermarking. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21050–21074, Suzhou, China. Association for Computational Linguistics.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. 2024a. A semantic invariant robust watermark for large language models. In *Proceedings of the 12th International Conference on Learning Representations*.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. 2024b. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57(2):1–36.

Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, and 1 others. 2024. Marker-llm: An open-source toolkit for llm watermarking. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 61–71.

Qi Pang, Shengyuan Hu, Wenting Zheng, and Virginia Smith. 2024. No free lunch in llm watermarking: Trade-offs in watermarking design choices. *Advances in Neural Information Processing Systems*, 37:138756–138788.

Jie Ren, Han Xu, Yiding Liu, Yingqian Cui, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. 2024. A robust semantics-based watermark for large language model

against paraphrasing. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 613–625.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2025. Can ai-generated text be reliably detected? stress testing ai text detectors under various attacks. *Transactions on Machine Learning Research*.

Shangqing Tu, Yuliang Sun, Yushi Bai, Jifan Yu, Lei Hou, and Juanzi Li. 2024. Waterbench: Towards holistic evaluation of watermarks for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1517–1542.

Qilong Wu and Varun Chandrasekaran. 2024. Bypassing llm watermarks with color-aware substitutions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8549–8581.

Yangxinyu Xie, Xiang Li, Tanwi Mallick, Weijie Su, and Ruixun Zhang. 2025. Debiasing watermarks for large language models via maximal coupling. *Journal of the American Statistical Association*, 120(551):1424–1436.

Hanlin Zhang, Benjamin L Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. 2024. Watermarks in the sand: impossibility of strong watermarking for language models. In *Forty-first International Conference on Machine Learning*.

Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu-Xiang Wang. 2024. Provable robust watermarking for AI-generated text. In *The Twelfth International Conference on Learning Representations*.

A CAP-Diag Diagnostic Details

Fano-style error-floor derivation. Let $W \in \{0, 1\}$ have prior π, let S be any possibly randomized function of the observed text, and let $\hat{W} = h(S)$ be any possibly randomized classifier with error $P_e = \Pr[\hat{W} \neq W]$. Fano's inequality gives

$$H(W \mid S) \leq H_b(P_e) + P_e \log_2(|\mathcal{W}| - 1).$$

For binary W, the second term is zero, and therefore

$$H(W) - I(W; S) \leq H_b(P_e).$$

Taking the inverse binary entropy on the branch $[0, 1/2]$ gives

$$P_e \geq H_b^{-1}(H(W) - I(W; S)).$$

The bound also applies to the Bayes-optimal classifier $\hat{W}^\star$, whose error $P_e^\star$ is the minimum over all classifiers consuming S. When $I(W; S) \to 0$, the bound approaches the prior-only baseline $P_e^\star \to \min(\pi, 1 - \pi)$.

Prior dependence. The main experiments use a balanced empirical prior $\pi = 0.5$. The mutual information $I(W; S)$ depends on the joint distribution of (W, S) and therefore on the prior. Under deployment prior shift to π', one should either reweight the empirical distributions to estimate $I_{\pi'}(W; S)$ or interpret the reported $I_{0.5}(W; S)$ as the balanced-detection diagnostic. For example, when $\pi' = 0.1$, $H(W) = H_b(0.1) \approx 0.469$ bit, which changes the numerical Fano floor but not the qualitative conclusion that low statistic-accessible MI limits any classifier consuming the same statistic.

Classifier lower bound. As a sensitivity check, we estimate a classifier lower bound on $I(W; S)$ using a cross-entropy objective on balanced samples ($n_{\text{wm}} = n_{\text{no}}$, $\pi = 0.5$). A logistic classifier $\hat{p}_\theta(W = 1 \mid S)$ is trained with 5-fold stratified cross-validation, and

$$\hat{I}_{\text{LB}} = 1 - \text{NLL}_{\text{CV}}(\hat{p}_\theta),$$

where NLL is measured in bits. The base feature is the detector score; the richer feature vector S' additionally includes token count, green count, and green rate for greenlist detectors.

Algorithm 1 CAP-Diag

Require: source corpus $\mathcal{D}$, rewrite regimes R, watermark family f, strengths Θ_f, utility gate $\mathcal{G}$, generator g, sample budget K

for each source x_i, regime r, family f, and strength θ **do**
  Draw K candidate rewrites from $g(\cdot \mid \pi_r(x_i))$ under watermark $\mathcal{W}_f(\theta)$
  Retain $\tilde{X}_{i,r,f,\theta} = \{x' : \mathcal{G}(x_i, x') = 1\}$
end for
for each regime r, family f, and strength θ **do**
  Compute the chosen detector statistic S on watermarked and unwatermarked accepted rewrites
  Estimate $\hat{I}(W; S)$ with clipped Miller–Madow histograms
  Compute source-paired bootstrap confidence intervals
  Compute AUROC and TPR at fixed FPR on the same balanced cell
end for
Run length, gate, estimator, and key-sensitivity diagnostics
Use the diagnosed low-information regime to derive attack and abstention readouts

Perplexity ratio. The PPL ratio used in Appendix E is the median per-source ratio of the watermarked generator's perplexity on a rewrite divided by the same generator's perplexity on the original source, both computed by Qwen2.5-7B-Instruct.

A.1 CAP-Diag Algorithm

A.2 Diagnostic Specification

Inputs. CAP-Diag takes as input: (i) a source set $\mathcal{D} = \{x_i\}$; (ii) a controlled rewrite ladder $R \in \{\text{R0}, \text{R1}, \text{R1.5}, \text{R2}\}$ defined by prompt and length band; (iii) a utility gate $\mathcal{G}$ enforcing non-contradiction, fact-skeleton preservation, and length-band membership; (iv) a watermark family f and strength θ; (v) a chosen detector statistic S; and (vi) an empirical prior π over $W \in \{0, 1\}$.

Procedure. For each source, regime, and watermark family, CAP-Diag samples accepted rewrites, computes the chosen detector statistic on matched watermarked and unwatermarked cells, estimates $\hat{I}(W; S)$ under the specified prior, and reports paired uncertainty. It then runs confound checks and uses the diagnosed low-information regime to derive the Low-Capacity Rewrite Attack and CAP-

Triage readout.

Outputs and scope. The output is a statistic-conditional accessible-information ladder, standard detection metrics on the same cells, robustness diagnostics, and downstream attack/readout consequences. The diagnostic is not a Shannon channel-capacity estimate and is not an all-detector upper bound. CAP-Triage is an abstention readout, not a recovery detector.

B Rewrite Regimes and Corpus Details

Rewrite regimes. R0 is a compressed fact-skeleton rewrite that preserves the key facts of the source while using a short, canonicalized form. R1 is an ordinary short paraphrase. R1.5 allows looser surface variation while remaining utility-preserving. R2 is a longer elaboration that retains the same source content while exposing a larger observed span to the detector. Full prompt templates and per-regime sample outputs are included in the supplementary material.

Synthetic corpus. The synthetic corpus contains 200 fact-skeleton episodes with controlled entity, value, and date fields. These episodes make fact preservation checkable while still allowing multiple surface rewrite regimes.

Naturalistic corpus. The naturalistic corpus contains 300 Wikipedia passages spanning fifteen domains. The eight frequent calibration domains used in Sec. 5.1 are *biography*, *arts/culture*, *sports*, *geography*, *science*, *general*, *history*, and *politics*. The seven held-out domains are *business*, *education*, *media*, *military*, *religion*, *technology*, and *transportation*. R0 and R1 are generated for the full naturalistic corpus; R1.5 and R2 are evaluated on a 150-source subset.

B.1 Cross-Generator Check

Llama-3.1-8B-Instruct on 60 synthetic sources at peak strength reproduces the R0 $\ll$ R1 ordering for KGW and UW. The higher regimes saturate earlier on this smaller validation run, so we treat the result as a cross-generator sanity check rather than a full robustness study.

C Utility Gate and Rewrite-Quality Validation

The utility gate is used to keep CAP-Diag focused on utility-preserving rewrites rather than semantic drift. It checks bidirectional non-contradiction, fact-skeleton preservation, and regime-specific length-band membership.

Post-gate accepted counts on the main sweep are $n = 400$ per cell for KGW and UW and $n = 200$ per cell for the EXP/Gumbel reduced sweep. R1.5 has the lowest pass rate because its looser prompt produces more NLI-rejected drafts.

Table 5: Utility-gate pass rates on a 50-source calibration subset with $K = 8$ drafts per source.

Regime	Pass rate	Rewrite type
R0	0.955	fact skeleton
R1	0.762	paraphrase
R1.5	0.297	loose rewrite
R2	0.830	elaboration

Table 6: Gate and skeleton-policy sensitivity. Values are R1/R0 MI ratios; the R0<R1 ordering is stable under these changes.

Check	KGW	UW
NLI thresholds $\{0.3, 0.5, 0.7\}$	5.19–5.60$\times$	7.35–7.40$\times$
Strict skeleton	8.34$\times$	13.65$\times$
Relaxed format	8.29$\times$	13.46$\times$
Soft entity-only	7.95$\times$	13.44$\times$

C.1 Gate-Threshold and Skeleton-Policy Sensitivity

The skeleton-policy ablation covers strict verbatim fact preservation, relaxed year/number formatting, and a soft entity-only policy. All three preserve the R0 versus R1 gap, indicating that the observed bottleneck is not induced by a single gate setting.

C.2 Independent LLM Audit

We sampled 500 accepted rewrites stratified across the four rewrite regimes and the Low-Capacity attack class. The audit model is DeepSeek v4-pro with deterministic decoding and a fixed prompt. This audit model is independent from the Qwen2.5-7B-Instruct generator used to produce the rewrites. A rewrite passes if it preserves key facts, introduces no factual drift, and preserves the main source meaning.

This audit is not a human evaluation. Its role is to provide an independent model-based quality check on accepted rewrites, separate from both the automatic utility gate and the generator used to produce the rewrites. The audit is applied after gate filtering, so the pass rates characterize the quality of the accepted evaluation cells rather than the acceptance rate of all candidate drafts.

Table 7: MI bin-count sensitivity at peak strength on the same R0/R1 cells as Table 2. The R0<R1 ordering is stable across bin counts.

Corpus	Regime	Family	10 bins	20 bins	30 bins	50 bins
Synthetic	R0	KGW	0.128	0.131	0.136	0.147
	R0	UW	0.131	0.136	0.127	0.170
	R0	EXP	0.063	0.074	0.078	0.101
	R1	KGW	0.726	0.882	0.865	0.886
	R1	UW	0.925	0.948	0.950	0.954
	R1	EXP	0.731	0.751	0.776	0.774
Naturalistic	R0	KGW	0.207	0.207	0.211	0.207
	R0	UW	0.125	0.136	0.145	0.128
	R0	EXP	0.056	0.058	0.063	0.063
	R1	KGW	0.858	0.858	0.861	0.854
	R1	UW	0.707	0.724	0.729	0.741
	R1	EXP	0.668	0.689	0.700	0.699

Table 8: Independent LLM audit on accepted rewrites.

Class	n	Passed	Pass rate
R0	100	99	99%
R1	100	100	100%
R1.5	100	100	100%
R2	100	100	100%
LowCap attack	100	99	99%

D Estimator and Detector Robustness

MI estimators. We evaluate whether the R0 versus R1 separation depends on the default clipped Miller–Madow histogram estimator. Equal-width histograms, equal-frequency histograms, classifier lower bounds on S, and classifier lower bounds on S' all preserve the R0$\ll$R1 ordering with disjoint uncertainty intervals in the main cells. The richer feature vector S' includes detector score, token count, green count, and green rate for greenlist detectors. The mixed continuous–binary kNN estimator saturates at one bit for these sample sizes, so we do not use it for fine-grained comparisons among high-information cells. This saturation does not affect the main low-information diagnosis, which is based on the stable R0 versus R1 separation.

Adaptive detector checks. We also test whether the R0 bottleneck disappears under stronger statistics derived from the same detector evidence. For KGW and UW, Higher Criticism, truncated goodness-of-fit, and Stouffer aggregation preserve the R0$\ll$R1 gap. For EXP/Gumbel, the continuous exponential per-token statistic makes Higher Criticism and truncated goodness-of-fit less informative, while Stouffer aggregation preserves the large R1/R0 ratio. These checks support the conclusion that the R0 bottleneck is not an artifact of a single thresholding rule.

Secret-key sensitivity. Three additional keys on KGW/UW preserve the R0 versus R1 gap. The observed ratios range from 3.3–$17.3\times$ for KGW and 5.9–$12.1\times$ for UW, with all confidence intervals disjoint.

D.1 Bin-Count Sensitivity

At 20, 30, and 50 bins, the estimates agree within approximately 0.05 bit in the main cells. At 10 bins, some R1 cells lose information because the broad score support is coarsely discretized, but the ordering and ratio-scale conclusion remain stable.

E CAP-Triage Reference Readouts and PPL Calibration

The main text focuses on the source-free length proxy because it is available without the original source or an auxiliary language model. We also evaluate oracle-regime and source-relative PPL readouts as reference diagnostics.

F Per-Domain Naturalistic Robustness

A per-family logistic readout $\hat{C}_r = \sigma(a_f H_{\mathrm{PPL}}(r) + b_f)$, with $H_{\mathrm{PPL}}(r) = -\log \overline{\mathrm{ppl_ratio}}_r$, descriptively tracks the regime means. With four regime points and two parameters, this should be read as interpolation over regimes, not as prediction for individual sources. Within a regime, the per-source PPL residual carries near-zero information about the z-gap in seven of eight regime/family cells.

Table 9: Per-domain naturalistic R0/R1 MI for KGW in domains with $n \geq 16$. Smaller domains have wide intervals and are treated as qualitative checks.

Domain	n	R0 MI	R1 MI	R1/R0
biography	78	0.18 [0.10, 0.34]	0.86 [0.77, 0.94]	4.78
arts/culture	43	0.21 [0.10, 0.42]	0.85 [0.71, 0.96]	4.05
sports	39	0.14 [0.07, 0.34]	0.93 [0.83, 1.00]	6.64
geography	34	0.22 [0.10, 0.45]	0.87 [0.74, 0.97]	3.95
science	24	0.32 [0.13, 0.60]	0.84 [0.65, 1.00]	2.63
general	18	0.15 [0.04, 0.45]	0.82 [0.60, 1.00]	5.47
history	16	0.27 [0.11, 0.61]	0.83 [0.59, 1.00]	3.07

Table 10: Held-out calibration of the regime-level PPL readout. The two-parameter logistic fit is descriptive rather than source-level predictive.

Held-out	Family	Measured	Reproduced	Error
R1.5	KGW	0.780	0.679	0.101
R1.5	UW	0.807	0.618	0.189
R2	KGW	0.995	1.000	0.005
R2	UW	0.984	1.000	0.016

Negative proxy result. The composite proxy

$$H_{\text{proxy}} = \log(c+1) + \rho - \log(\text{ppl_ratio}),$$

where c is the NLI-passing rewrite count among $K = 8$ samples, fails because R1.5 has the lowest accepted-draft count. Plugging the observed R1.5 proxy into the logistic readout predicts R1.5 MI around 0.04 against a measured value around 0.78.

G Low-Capacity Rewrite Attack Details

The EXP/Gumbel pass of the Low-Capacity Rewrite Attack mirrors the KGW and UW passes: 100 source episodes, the same R0 prompt, the same utility gate, and the same MI estimator with source-paired bootstrap. Pre-attack documents are EXP-watermarked R2 long-form rewrites; the attack applies an R0 prompt re-rewrite with no watermark. The detector is the EXP/Gumbel statistic

$$S_T = \sum_t -\log(1 - r_{x_t}).$$

Pre-attack MI is 1.000 with interval $[0.977, 1.000]$; post-attack MI is 0.093 with interval $[0.081, 0.241]$; the MI drop is 90.7% with interval $[77.2, 93.0]\%$. Pre-attack AUROC is 1.000, and post-attack AUROC is 0.663. The rounded value reported in the main text is 0.64 in the figure/table readout.

H Mechanism Reference Details

Sentence-level semantic reference. The sentence-level semantic reference embeds candidate sentences, assigns them to secret-keyed semantic buckets, and detects with a per-text green-bucket statistic. This reference is used only to test whether a non-token-count statistic exhibits the same R0 versus R1 information gap. It is not an exhaustive evaluation of semantic or structure-aware watermarking methods.

Content-presence reference. We use the all-MiniLM-L6-v2 SBERT model to embed source passages and a fixed seed list of $4{,}096$ content tokens. Per source, the inserter selects the top-$K = 8$ nearest-neighbor seeds by cosine similarity in SBERT space, then inserts them into the unwatermarked output at approximately one insertion per 12 tokens, with positions randomized under a secret-keyed RNG. Survival of inserted seeds under rewrite is simulated at rate

$$p_{\text{survive}} = \text{overlap}(\text{source}, \text{rewrite}) \times (1 - 0.55),$$

where the penalty captures the empirical tendency of rare insert tokens to be dropped by meaning-preserving paraphrasers. The detector predicts the same top-K seeds from the source and counts predicted-seed presence in the candidate text.

This reference follows the content-insertion direction of PostMark (Chang et al., 2024b), but it is not a reproduction of the full PostMark pipeline. We use it only as a mechanism contrast. Its R1/R0 ratio remains close to one on both corpora, as expected for a content-presence statistic that depends on source-token survival rather than token-distribution deviation.

The seven domains in Table 9 all show R1/R0 ratios above one. Domains with fewer than 16 paired observations produce wide intervals and are reported only in the supplementary material.

I Reproducibility Details

The canonical secret key is 20260513; the additional keys used for key-sensitivity checks are 13371337, 99999991, and 42424242. All bootstrap procedures use the same canonical resampling seed. Generation with Qwen2.5-7B-Instruct uses top-$p = 0.95$ sampling, temperature 0.7, a 160-token cap for R0 documents, and longer caps for R1/R1.5/R2. The NLI gate uses the `cross-encoder/nli-deberta-v3-base` model with a bidirectional contradiction threshold of 0.5.

Detection uses vocabulary size $V = 152{,}064$ for KGW and UW with green-list fraction $\gamma = 0.25$. EXP/Gumbel uses a context window of four tokens and the per-token statistic described in Sec. 2. Source identifiers, prompt templates, and aggregate detector scores are included in the supplementary material.

J Reproducibility and Compute Details

Compute infrastructure. The local experiments were run on a single-workstation setup with one NVIDIA GeForce RTX 5090 GPU. The GPU reports 32,607 MiB of memory under NVIDIA-SMI, with NVIDIA driver version 580.97 and CUDA version 13.0. The host CPU is an Intel Core i9-14900KF with 24 physical cores and 32 logical processors, and the machine has approximately 32 GB of system memory. All local GPU-based runs used a single GPU.

Models. The primary rewrite generator is Qwen2.5-7B-Instruct. The cross-generator validation uses Llama-3.1-8B-Instruct. The independent rewrite-quality audit uses DeepSeek v4-pro and is therefore reported separately from the local GPU-based experiments. The detector-side experiments evaluate KGW, UW, and EXP/Gumbel watermark statistics, as described in Sec. 2 and Sec. 3.2.

Experimental configuration. The main CAP-Diag experiments use balanced watermarked and unwatermarked evaluation cells, a balanced prior $\pi_{\mathrm{wm}} = 0.5$, Miller–Madow-corrected histogram mutual information estimates, source-paired bootstrap intervals, AUROC, and TPR at fixed false-positive rate. The controlled rewrite ladder contains R0, R1, R1.5, and R2, and accepted rewrites must pass the utility gate described in Appendix C. The naturalistic corpus contains 300 Wikipedia passages across fifteen domains; R0 and R1 are generated for the full naturalistic corpus, while R1.5 and R2 are evaluated on a 150-source subset.

Compute budget accounting. Exact aggregate GPU-hours were not instrumented in the experiment logs. We therefore report the hardware configuration above and the sample counts in Sec. 3.2, Tables 2 and 3, and the appendices, rather than claiming an exact GPU-hour total. The LLM audit is API-based and is not counted as local GPU time.