

NeurIPS 2026 Main Track | Paper #36189

A Retained-Signal Interface for LLM Watermark Robustness under Paraphrase

Chen Wang^{1,2} Yinxuan Huang³ Yexin Cui³ Maoqing Zhong³
LaiLong Luo^{1,2*}

¹College of Systems Engineering, National University of Defense Technology, Changsha 410073, Hunan, China

²National Key Laboratory of Information Systems Engineering, NUDT, Changsha 410073, Hunan, China

³College of Computer Science, National University of Defense Technology, Changsha 410073, Hunan,

Core Takeaway

Post-paraphrase watermark detectability is governed by the **retained-signal product** $\Delta\rho$, not Δ alone.

Small-sample score calibration ($m=20-50$) predicts full-run AUC with **MAE 0.016**.

One-Sentence Summary: If a watermark injects KL signal $\Delta = \text{KL}(P_1||P_0)$ and a paraphrase channel retains fraction ρ , post-paraphrase detectability is governed by $\Delta\rho$, enabling a calibrated small-sample score audit $\widehat{\text{AUC}} = \Phi(\sqrt{\widehat{\Delta\rho}})$ and a reproducible four-part robustness report.

1. Why Replace One-Off “AUC Under Paraphraser X” Tables?

Standard single-number robustness evaluation conflates four distinct mechanisms

NeurIPS 2026

Paper #36189 | Slide 2/10

Four Conflated Factors in Standard Watermark Evaluation

Reporting a single post-paraphrase AUC (e.g., “AUC = 0.72 after DIPPER”) mixes four separate effects:

- **1. Injected Signal $\Delta = \text{KL}(P_1 \| P_0)$:**
How strongly the watermark perturbs token sampling before any rewrite (scales with bias δ and window length L).
- **2. Channel Contraction ρ :**
What fraction of the injected KL evidence survives the paraphrase channel W .
- **3. Semantic Preservation:**
Whether the rewrite kept meaning intact (embedding cosine similarity, length ratio).
- **4. Detector Projection $s(X)$:**
Which 1D test statistic (e.g., green-list z -score vs. learned classifier) is actually being evaluated.

Diagnostic Goal: Separate pre-paraphrase injection (Δ) from channel survival (ρ) and semantic drift.

Core Research Questions & Running Benchmarks

Q1 (Global Converse): What full-distribution unit controls *every* detector after paraphrase?

→ **The retained-signal product $\Delta\rho$ at sharp rate $\Theta(\sqrt{\Delta\rho})$.**

Q2 (Implementation-Blindness): At matched Δ , do cryptographic and public statistical watermarks share the same converse?

→ **Yes:** $B^*(\Delta, \varepsilon) = \frac{1}{2} \log(\Delta/2\varepsilon^2)$.

Q3 (Practical Audit): Can $\Delta\rho$ be estimated from small calibration samples for a deployed score $s(X)$?

→ **Yes: score-projected plug-in $\hat{\Delta}\hat{\rho}$ predicts AUC with MAE 0.016.**

Three Watermark Families Evaluated:

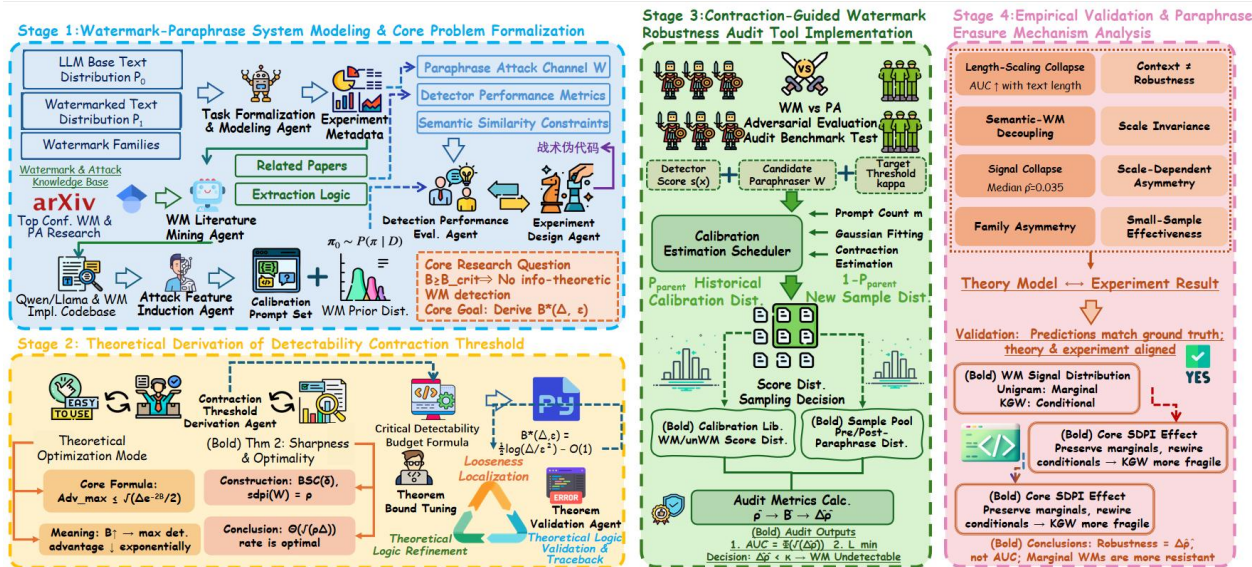
- **KGW:** Context-hashed green list (γ, δ); green-fraction z -score.
- **Unigram:** Fixed global green list; marginal token shift; same z -score.
- **EXP / Gumbel-max:** Keyed sampling
 $\arg \max_v \{\log p_v - \log(-\log u_v)\}$; detected via $\sum_t -\log(1 - u_t)$.

2. Retained-Signal Interface Overview

Four-stage workflow connecting information-theoretic contraction to empirical score auditing

NeurIPS 2026
Paper #36189 | Slide 3/10

End-to-End Architecture: Modeling → Converse Theory → Score Auditor → Empirical Collapse



$$\underbrace{\Delta = \text{KL}(P_1 \| P_0)}_{\text{Pre-paraphrase injected signal}} \times \underbrace{\rho = e^{-2B}}_{\text{Channel retained fraction}} = \underbrace{\Delta\rho}_{\text{Effective retained signal}} \Rightarrow \widehat{\text{AUC}} = \Phi\left(\sqrt{\Delta\rho}\right)$$

3. Problem Formulation & The Three-Level Reader Contract

Strictly separating what is proved globally from what is audited on a detector score

NeurIPS 2026
Paper #36189 | Slide 4/10

Binary Testing Across Paraphrase Channel W

- Pre-paraphrase distributions:** Base LLM P_0 vs. watermarked LLM P_1 with window KL signal

$$\Delta := \text{KL}(P_1 \| P_0).$$

- Paraphrase channel W :** Contracts KL separation with SDPI coefficient and log-budget

$$\text{sdpi}(W) = \sup_{P \neq Q} \frac{\text{KL}(WP \| WQ)}{\text{KL}(P \| Q)} \leq \rho = e^{-2B}.$$

- Optimal testing advantage:**

$$\sup_{\phi} \text{Adv}(\phi; WP_0, WP_1) = \text{TV}(WP_0, WP_1),$$

and for any ROC curve, $\text{AUC} - \frac{1}{2} \leq \text{TV}(WP_0, WP_1)$.

Data-Processing Bottleneck: No post-processing detector can exceed the channel-contracted TV distance.

Three Levels: Proved vs. Measured (Reader Contract)

Level	Symbols	Formal Meaning & Scope
1. Global (Theory)	$\Delta = \text{KL}(P_1 \ P_0)$ $\rho = \text{sdpi}(W)$ $B = -\frac{1}{2} \log \rho$	Property of the full text pair (P_0, P_1) and channel W . Drives the detector-independent converse (Theorems 1, 2, 7).
2. Score-Projected (Audit Target)	$\Delta_s = \text{KL}(S_1 \ S_0)$ $\rho_s = \frac{\text{KL}(\tilde{S}_1 \ \tilde{S}_0)}{\Delta_s}$	Same quantities after projecting text through a named score $s: \mathcal{X} \rightarrow \mathbb{R}$ ($S_i = s_{\#} P_i, \tilde{S}_i = s_{\#} W P_i$). Proposition 3.
3. Plug-In (Reported)	$\hat{\Delta}, \hat{\rho}, \hat{\Delta}\hat{\rho}$	Calibration-sample estimates of (Δ_s, ρ_s) . Every empirical number in the paper lives at this level.

Scope Rule: “Detector-independent” refers exclusively to Level 1. Every empirical $\hat{\rho}$, predicted AUC, and calibration error applies to the evaluated score projection $s(X)$.

4. Main Theoretical Results: Converse, Scaling & Sharpness

Pinsker/Bretagnolle–Huber envelopes, length scaling, BSC rate optimality, and implementation-blindness

NeurIPS 2026

Paper #36189 | Slide 5/10

Contraction Converse (Theorems 1 & 2, Proposition 3)

For any (P_0, P_1) with $\text{KL}(P_1 \| P_0) \leq \Delta$ and channel W with contraction $\rho = e^{-2B}$:

$$\sup_{\phi} \text{Adv}(\phi) \leq \min\left\{\sqrt{\frac{\Delta\rho}{2}}, \sqrt{1 - e^{-\Delta\rho}}\right\} \leq 1 - \frac{1}{2}e^{-\Delta\rho}.$$

To force every detector's advantage below $\varepsilon \in (0, 1)$, it suffices that

$$B \geq B^*(\Delta, \varepsilon) := \frac{1}{2} \log\left(\frac{\Delta}{2\varepsilon^2}\right).$$

Score-Projected Converse (Proposition 3): For any detector family

$\mathcal{D}_s = \{\phi = g \circ s\}$ thresholding or post-processing score $s(X)$, advantage satisfies $\sup_{\phi \in \mathcal{D}_s} \text{Adv}(\phi; WP_0, WP_1) \leq \sqrt{\Delta_s \rho_s / 2}$.

Accumulation, Rate Optimality & Implementation-Blindness

- Window Accumulation & Minimum Length (Corollaries 4–5):**

With additive signal $\Delta_L \approx L\bar{\Delta}$, reaching advantage ε requires

$$L \geq L_{\min}(\bar{\Delta}, \rho, \varepsilon) := \frac{2\varepsilon^2}{\bar{\Delta}\rho}.$$

- Serial Paraphrase Composition (Corollary 6):**

Under $W_m \cdots W_1$, contractions multiply ($\rho_{\text{tot}} \leq \prod_j \rho_j$) and budgets add ($B_{\text{tot}} \geq \sum_j B_j$).

- (Δ, ρ) -Only Sharpness Witness (Theorem 7 & Corollary 8):**

On $W = \text{BSC}(\delta)$ with $(1 - 2\delta)^2 = \rho$ and $P_0 = \text{Ber}(1/2)$, $P_1 = \text{Ber}(1/2 + \theta)$,

$$\text{TV}(WP_0, WP_1) = \sqrt{\frac{\rho\Delta}{2}} (1 + o(1)),$$

proving $\Theta(\sqrt{\Delta\rho})$ cannot be improved by any (Δ, ρ) -only bound.

- Implementation-Blindness at Matched Δ (Corollary 9):**

Cryptographic (PRF) and public statistical watermarks with the same Δ share the exact same distributional converse $B^*(\Delta, \varepsilon)$.

Unified Scaling Law: Both single-pass and recursive paraphrasing obey the same $\Delta\rho$ bottleneck.

5. Algorithm A1: Contraction-Guided Score Auditor

Calibrating post-paraphrase AUC and minimum detection length from $m \in \{10, 20, 50\}$ prompts

NeurIPS 2026

Paper #36189 | Slide 6/10

8-Step Score-Projected Audit Protocol

Input: Deployed score $s(\cdot)$, paraphraser W , m calibration prompts $\mathcal{C}$, length L , threshold κ , floor $\eta = 10^{-9}$.

1. **Paired Generation:** Draw clean base $x_0(c) \sim P_0$ and watermarked $x_1(c) \sim P_1$ for each $c \in \mathcal{C}$.
2. **Paraphrase:** Obtain $\tilde{x}_0(c) = W(x_0(c))$ and $\tilde{x}_1(c) = W(x_1(c))$.
3. **Score Projection:** Compute 1D score sets $S_0, S_1, \tilde{S}_0, \tilde{S}_1$.
4. **Gaussian / Histogram Fit:** Fit 1D laws $\hat{S}_0, \hat{S}_1, \hat{\tilde{S}}_0, \hat{\tilde{S}}_1$.
5. **Signal & Contraction:**

$$\hat{\Delta} = \text{KL}(\hat{S}_1 \| \hat{S}_0), \quad \hat{\rho} = \frac{\text{KL}(\hat{\tilde{S}}_1 \| \hat{\tilde{S}}_0)}{\max\{\hat{\Delta}, \eta\}}.$$

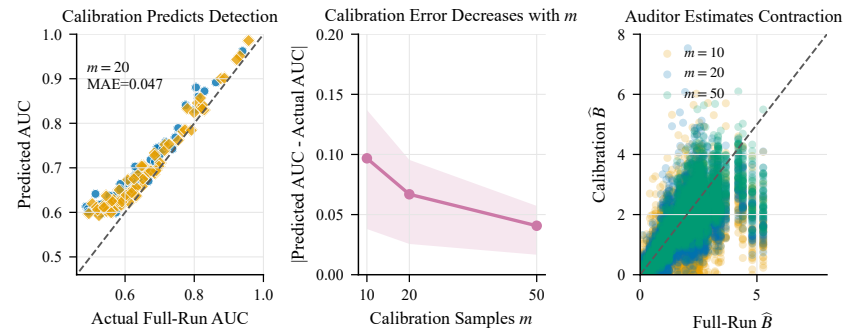
6. **Plug-In AUC Prediction:**

$$\hat{B} = -\frac{1}{2} \log \hat{\rho}, \quad \widehat{\text{AUC}} = \Phi(\sqrt{\hat{\Delta} \hat{\rho}}).$$

7. **Length Extrapolation:** $\hat{L}_{\min} = \lceil L\kappa / (\hat{\Delta} \hat{\rho}) \rceil$.

8. **Decision:** Return BELOWTHRESHOLD if $\hat{\Delta} \hat{\rho} < \kappa$.

Small-Sample Auditor Validation (Finding 4)



Calibration accuracy vs. prompt count m :

- $m = 10$ prompts: full-run AUC MAE = 0.097 (correlation 0.63)
- $m = 20$ prompts: full-run AUC MAE = 0.067 (correlation 0.79)
- $m = 50$ prompts: full-run AUC MAE = 0.041 (correlation 0.92)

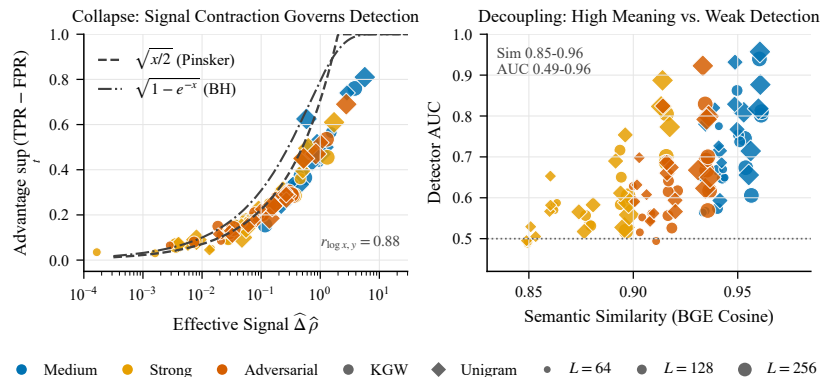
Estimator robustness (144 runs): Nonparametric histogram KL and Gaussian plug-in agree on $\log_{10}(\hat{\Delta} \hat{\rho})$ with $r = 0.91$ (Gaussian AUC MAE 0.016 vs. histogram 0.031).

6. Empirical Collapse & Semantic Decoupling (108-Cell Sweep)

KGW/Unigram $\times$ Qwen/Llama paraphrasers $\times L \in \{64, 128, 256\} \times 3$ strengths $\times 3$ (γ, δ) settings

NeurIPS 2026
Paper #36189 | Slide 7/10

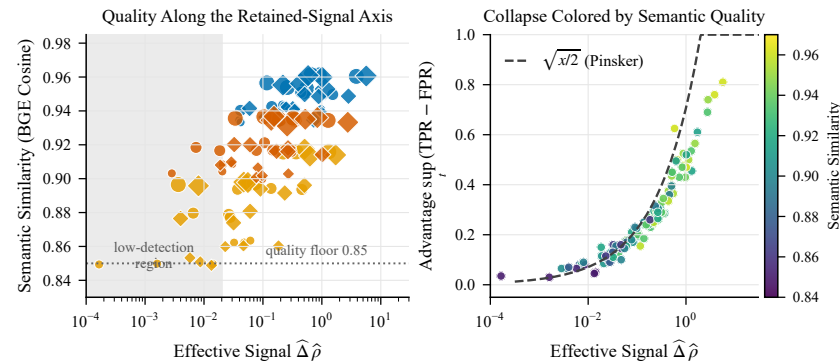
Finding 1: Effective-Signal Collapse onto Envelopes



- **Tight collapse:** Spearman correlation with $\sup_t(\text{TPR} - \text{FPR})$ is **0.983** for $\log(\hat{\Delta}\hat{\rho})$, vs. **0.104** for $\log \hat{\Delta}$ alone.
- **Envelope compliance:** 107/108 cells lie below the histogram-KL Pinsker/BH envelope (max excess 0.001, inside Monte-Carlo band 0.122).
- **Plug-in calibration:** $\Phi(\sqrt{\hat{\Delta}\hat{\rho}})$ predicts deployed AUC with **MAE 0.016**.

Key Insight: Pre-paraphrase $\hat{\Delta}$ alone is uninformative ($\rho_S = 0.104$); multiplying by $\hat{\rho}$ unifies all 108 cells.

Finding 2 & 3: Low Detection Without Semantic Collapse



- **High meaning inside low-detection band:** Among 15 cells with $\hat{\Delta}\hat{\rho} \leq 2\epsilon^2$ ($\epsilon = 0.1$), 11/15 keep BGE similarity ≥ 0.85 (median 0.88).
- **Context vs. marginal sensitivity (Finding 3):** Fixed-list Unigram retains higher average AUC than context-hashed KGW (0.747 vs. 0.698 at $L = 256$) because paraphrase disrupts local token windows more than marginal vocabulary; the gap closes at matched $\hat{\Delta}$.

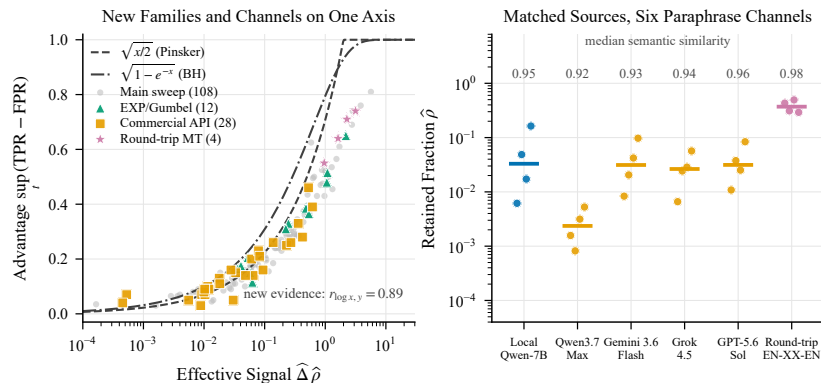
Decoupling: Paraphrasers can drive $\hat{\Delta}\hat{\rho}$ below the detection threshold while preserving semantic similarity ≥ 0.88 .

7. Generalization Across Watermark Families, APIs & Generators

Testing the retained-signal interface without family-specific recalibration (Finding 5)

NeurIPS 2026
Paper #36189 | Slide 8/10

Unified Collapse Across New Families & Channels



Zero Recalibration: All new watermark families, commercial APIs, round-trip translation channels, and source generators collapse onto the exact same $\hat{\Delta p}$ curve using the unchanged Algorithm A1 protocol.

Four Extended Generalization Audits

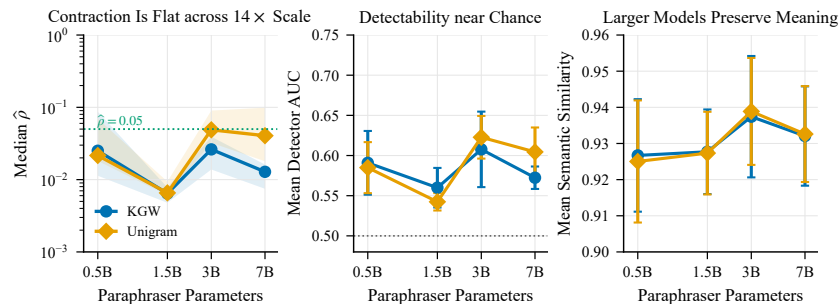
- 1. EXP / Gumbel-Max Family (12 cells):**
Keyed sampling detected via $\sum_t -\log(1 - u_t)$. Unchanged Gaussian plug-in predicts post-paraphrase AUC with **MAE 0.021** (median $\hat{p} = 0.021$, vs. KGW 0.022, Unigram 0.094).
- 2. Four Commercial API Paraphrasers (28 conditions):**
Qwen3.7-Max ($\hat{p} = 0.0028$), Grok-4.5 (0.0054), Gemini-3.6-Flash (0.0084), and GPT-5.6-Sol (0.0314). AUC prediction **MAE 0.034**; 12/28 enter low-detection at similarity > 0.85 .
- 3. Round-Trip Machine Translation (EN $\rightarrow$ ZH/FR $\rightarrow$ EN):**
Reaches 0.98 semantic similarity yet retains $\sim 56\times$ more signal ($\hat{p} = 0.37$) than matched direct API paraphrase, keeping AUC at 0.83–0.93.
- 4. Second Source Generator (Llama-3.1-8B-Instruct):**
Holding prompts and paraphraser fixed, $\Phi(\sqrt{\hat{\Delta p}})$ tracks post-paraphrase AUC with **MAE 0.021** (max 0.039) at similarity 0.924–0.954.

8. Paraphraser Scale Ablation & Detector-Transfer Stress Test

Larger paraphrasers improve meaning preservation; simple learned projections do not recover lost signal

NeurIPS 2026
Paper #36189 | Slide 9/10

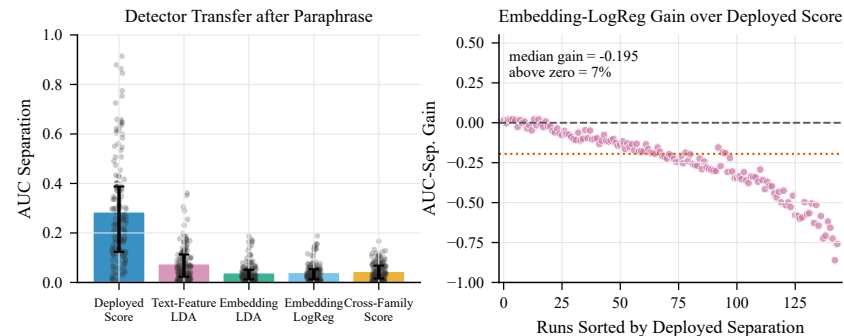
Paraphraser Scale Ablation (36 Runs)



Takeaway: Across paraphraser model sizes (0.5B to 7B), median score-projected $\hat{\rho}$ and post-paraphrase AUC remain nearly flat while embedding cosine similarity rises steadily.

→ **Scaling up the paraphraser mainly improves semantic fidelity within the same score-contraction regime.**

5-Fold Prompt-Level Detector Transfer



Median post-paraphrase $|AUC - 0.5|$ across held-out folds:

- Deployed watermark score z : **0.236**
- Surface-feature LDA: 0.057
- BGE-embedding LDA: 0.026
- BGE-embedding + LogReg: 0.032

Standard feature/embedding classifiers do not recover signal missed by z ; any new hybrid detector should be audited via its own projection $\hat{\Delta}_s \hat{\rho}_s$.

9. Conclusion: The Reproducible Four-Part Robustness Report

Replacing one-off post-attack AUC numbers with a calibrated retained-signal interface

NeurIPS 2026
Paper #36189 | Slide 10/10

Recommended Four-Part Watermark Robustness Report

Instead of reporting a single “AUC under rewriter X”, every watermark robustness evaluation should report four coupled quantities on the same outputs:

- 1. Pre-Paraphrase Signal $\hat{\Delta}$:**
Clean watermark KL separation before paraphrase (separates weak injection from strong attack).
- 2. Retained Signal $(\hat{\rho}, \hat{\Delta}\hat{\rho})$:**
Channel survival fraction $\hat{\rho}$ and effective retained signal $\hat{\Delta}\hat{\rho}$ (governs post-attack detectability and $L_{\min}$).
- 3. Semantic Preservation:**
Embedding cosine similarity and length ratio on the exact same paraphrased outputs.
- 4. Detector Projection $s(X)$:**
Explicit specification of the score statistic audited by $(\hat{\Delta}, \hat{\rho})$.

Explicit Scope Boundaries & Summary

Scope Boundaries:

- **Global vs. Projected:** Only Theorems 1–2 are detector-independent; all empirical numbers apply to the evaluated score projection $s(X)$.
- **Score Calibration:** Heavy-tailed or multimodal scores should use the nonparametric histogram KL option in Algorithm A1.
- **Open Directions:** Semantic watermarks, alignment-cost/hybrid detectors, and adaptive detector-aware attacks enter by auditing their own score projections.

Thank You!

Paper #36189 | Code & Audit Records in Supplement