REAGG: RELIABILITY-AWARE AGGREGATION OF TOKEN EVIDENCE FOR MACHINE-GENERATED TEXT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The widespread use of large language models makes machine-generated text detection crucial for content provenance. Existing detectors typically aggregate token-level statistics into document scores via fixed rules, such as simple means or ratios. However, such uniform aggregation treats every token equally, regardless of its position in the text or how consistently the two reference models represent it. We derive an exact variance excess identity that quantifies the cost of weighting mismatch and support loss, and use this result to guide the construction of reliability weights from two observable cues. Specifically, we propose ReAgg, a reliability-aware aggregation framework that computes a contribution weight for each token using two computable cues: (i) its position in the text, and (ii) the representational stability between a base model and its instruction-tuned counterpart. It further adapts these weights to four scoring interfaces—means, sums, ratios, and spectral statistics—without training any additional model. We evaluate ReAgg across eight carriers spanning five benchmark families (18 subsets), improving macro-average direction-corrected AUROC from 75.22% to 77.54% (+2.32 pp) across the 13 primary subsets. Ablations support the empirical complementarity of the position and hidden-discrepancy cues.

1 INTRODUCTION

The widespread deployment of large language models makes machine-generated text detection crucial for content provenance, academic integrity, and information authenticity (Dugan et al., 2024; Yu et al., 2026). When texts lack reliable provenance metadata and generating models remain inaccessible, detectors must evaluate texts post hoc. Statistical detection based on frozen reference language models provides an effective avenue without training additional detection networks: it extracts discriminative signals from the model's token-level response to form a document-level score (Su et al., 2023; Bao et al., 2024). For token-level methods, detection efficacy fundamentally depends on two decoupled components: how local evidence is constructed and how such evidence is aggregated into a document-level decision.

Existing zero-shot detectors extract local evidence z_i from token likelihoods, ranks, conditional probability curvature, and prediction divergence between paired reference models, yielding established baselines such as DetectLLM (Su et al., 2023), Fast-DetectGPT (Bao et al., 2024), and Binoculars (Hans et al., 2024). Despite diverse evidence formulations, many statistical carriers use fixed aggregation rules that do not explicitly adjust token contributions using relative position and base–instruction hidden-state discrepancy. Although recent studies explore non-uniform weighting, they either learn parameterized weights requiring supervised training (Miralles-González et al., 2025) or couple weight estimation with specific scoring formulations (Zhu et al., 2026). A general, training-free method that uniformly adjusts token weights across different types of detectors remains missing.

Local evidence magnitude and its contribution to a document score are two distinct concepts. As illustrated in Figure 1, tokens 2 and 5 share identical evidence values ($z_2 = z_5 = 0.60$). Under uniform averaging with $\alpha_i = 1/6$, both tokens receive the same contribution ($\alpha_2 z_2 = \alpha_5 z_5 = 0.100$), as shown in Figure 1(a). However, identical evidence does not mean these tokens should contribute equally—their positions and representational stability differ. ReAgg therefore constructs

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

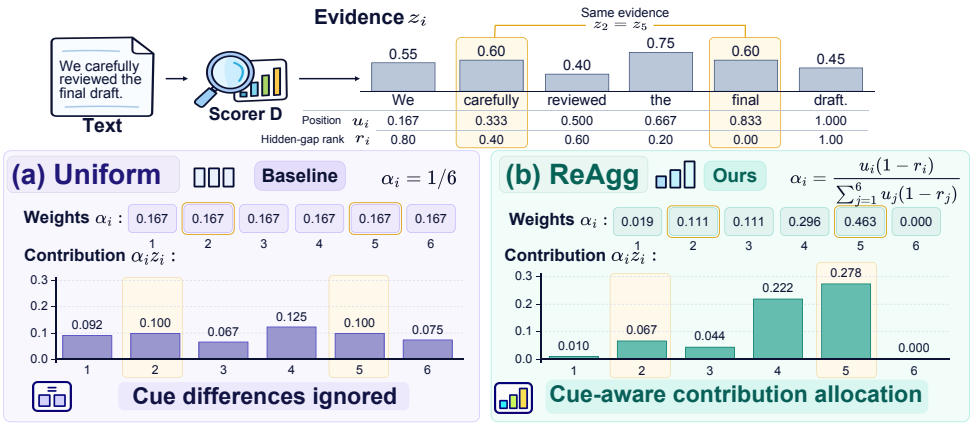


Figure 1: Conceptual comparison of token evidence aggregation. Given identical evidence ($z_2 = z_5 = 0.60$), **(a)** uniform averaging assigns equal contributions to the two highlighted tokens ($\alpha_2 z_2 = 0.100$), whereas **(b)** ReAgg recalibrates contributions via sequence position and hidden-state discrepancy ($\alpha_2 z_2 = 0.067$ vs. $\alpha_5 z_5 = 0.278$), leaving the evidence unchanged.

weights from two computable cues: (i) relative sequence position, which favors later positions; and (ii) hidden-state discrepancy between base and instruction-tuned models, which favors positions with smaller representational change. As shown in Figure 1(b), this rule assigns different contributions to tokens 2 and 5 ($\alpha_2 z_2 = 0.067$ vs. $\alpha_5 z_5 = 0.278$) while keeping the evidence values unchanged.

To address this challenge, we propose **ReAgg** (**Re**liability-Aware **Agg**regation), a training-free framework that decouples local evidence generation from document-level contribution assignment. To ground this design, we analyze evidence aggregation under a conditional variance model and derive an exact variance excess identity, showing how weighting mismatch and support loss affect aggregation quality (Section 3.2). Guided by this perspective, ReAgg constructs joint reliability weights $w_i = p_i \eta_i$ from two computable surrogates: a polynomial position prior p_i and a rank-normalized hidden-state discrepancy factor η_i between paired frozen models. To accommodate heterogeneous scoring operations across detector carriers, ReAgg establishes tailored aggregation operators for means, sums, ratios, and spectral statistics, producing reweighted document scores $S_{\text{cal}}^{(D)}$ without additional model training or per-dataset hyperparameter search.

We evaluate ReAgg across eight detector carriers on 18 subsets from five benchmark families, totaling 24,210 text records. Under a unified frozen protocol ($\beta = \gamma = 1$) without per-dataset tuning, ReAgg increases macro-average direction-corrected AUROC from 75.22% to 77.54% across the 13 primary subsets, with positive aggregate gains across all four scoring structures and five benchmark families. Systematic ablations and permutation controls confirm the empirical complementarity of position and representation cues, while grounded case studies delineate the operating boundaries where non-uniform aggregation succeeds or encounters limits.

Our main contributions are summarized as follows:

- **Variance theory for the weight-design stage.** We derive the exact identity $V(\mathbf{a}) - V^\star = V^\star \chi^2(\mathbf{a} \| \mathbf{a}^\star)$: the variance cost of any weighting scheme is its χ^2 distance from the oracle allocation. This identity quantifies the variance cost of weighting mismatch and support loss, providing the design criterion behind ReAgg's choice of computable surrogates.
- **Training-free reliability-aware framework.** We propose ReAgg, which decouples evidence extraction from contribution assignment. Combining relative position priors with base–instruction hidden discrepancies, ReAgg establishes closed-form operators across four scoring structures (means, sums, ratios, spectral statistics) without model training.
- **Systematic cross-benchmark evaluation.** Across 24,210 texts spanning five benchmarks and eight carriers, we systematically demonstrate the empirical efficacy of ReAgg under a unified protocol, while providing transparent analysis of failure modes and empirical boundary conditions.

2 Related Work

2.1 Supervised and Learning-Based Detection

Supervised detectors typically learn decision boundaries from labeled corpora containing human-written and machine-generated texts. Early approaches directly fine-tuned pretrained models such as RoBERTa (Solaiman et al., 2019) for text attribution; RADAR (Hu et al., 2023) incorporates adversarial training between a detector and a paraphraser to improve robustness against text rewriting; and Ghostbuster (Verma et al., 2023) trains classifiers on structured probability features extracted from multiple weaker language models. MAGE (Li et al., 2024) further establishes a cross-domain, multi-generator benchmark, demonstrating that the transfer capability of fine-tuned detectors requires empirical scrutiny under varying data distributions. For settings where the generation pipeline is unobservable, this work focuses on post-hoc statistical detection without task-specific classifier training, directly constructing detection scores from publicly available frozen reference models.

2.2 Reference-Model-Based Statistical Detection

Reference language models provide diverse statistical signals for machine-generated text detection via token probabilities, ranks, and cross-model prediction discrepancies. GLTR (Gehrmann et al., 2019) inspects token probability distributions and ranks; DetectLLM (Su et al., 2023) constructs detection metrics like LRR and NPR using log-rank information; DetectGPT (Mitchell et al., 2023) and Fast-DetectGPT (Bao et al., 2024) estimate local probability curvature through text perturbation and conditional sampling, respectively; Binoculars (Hans et al., 2024) measures the ratio of perplexity to cross-perplexity across model pairs; and IRM (Liu et al., 2025) uses the log-likelihood difference between instruction-tuned and base models for detection. Recent studies further examine sequence structures: Lastde (Xu et al., 2025) extracts local temporal dynamics from token probability sequences, while SpecDetect (Luo et al., 2026) performs frequency-domain spectral analysis directly on log-probability series.

Although these methods formulate diverse local statistics, many existing carriers rely on fixed aggregation rules—such as arithmetic means, unweighted sums, or fixed ratios—that overlook variations in sequence position and hidden-state discrepancy, which can allow noisy local evidence to affect the final score. To address this persistent aggregation bottleneck, this work strictly preserves underlying evidence definitions while adaptively recalibrating token contributions, establishing a plug-and-play enhancement across established detector carriers.

2.3 Token Reweighting and Detection Enhancement

Prior efforts have explored detection score enhancement through complementary mechanisms. For external-channel enhancement, TOCSIN (Ma & Wang, 2024) measures semantic variations across randomly perturbed token deletions to derive a cohesiveness score, subsequently fusing it with the base detector's document-level output. PAWN (Miralles-González et al., 2025) learns token weights from hidden states and positional information through supervised training. HAPDA (Jiang et al., 2026) fine-tunes paired human and machine preference models to adjust token contributions for mean-based detectors using prediction divergence and uncertainty. Exons-Detect (Zhu et al., 2026) amplifies tokens with high hidden-state discrepancies between paired models within its customized translation score. Similarly, Alignment Imprint's LAPD (Wu et al., 2026) applies self-information weighting to cross-model probability differences.

ReAgg constructs token weights directly from frozen base and instruction-tuned models, combining a relative-position prior with a low-discrepancy factor, and applies the same fixed rule through mean, sum, ratio, and spectral operators while retaining each carrier's original local evidence.

3 The ReAgg Framework

3.1 Problem Formulation and Framework Overview

Given a token sequence $x = (t_1, \ldots, t_T)$ of length T, the goal is to predict its provenance label $y \in \{0, 1\}$, where 1 denotes machine-generated text and 0 denotes human-authored text. For a detector

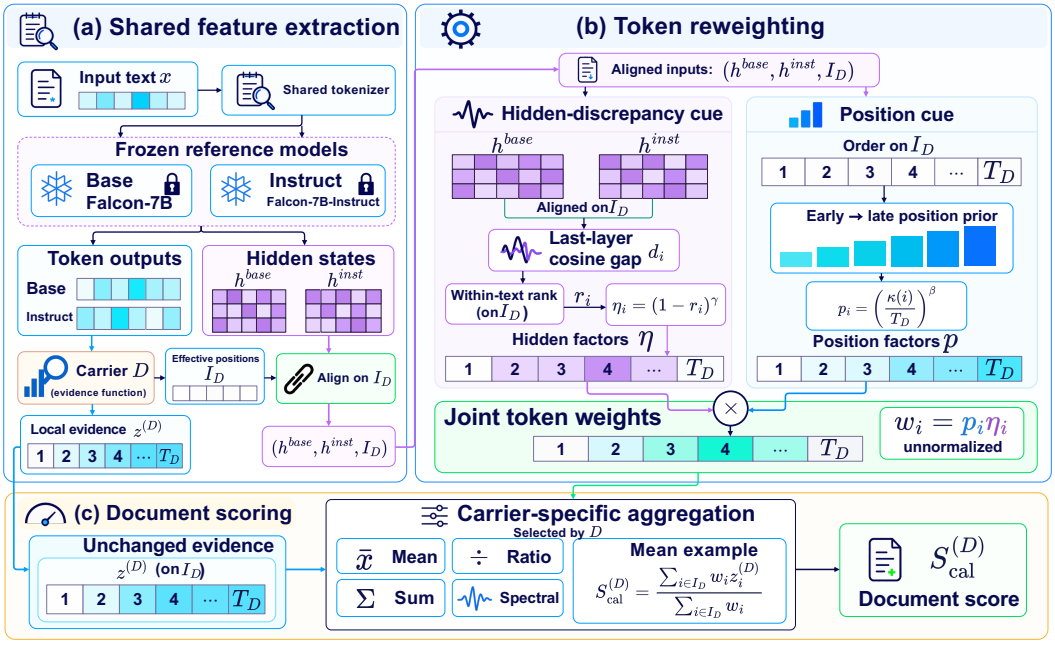


Figure 2: Overview of the ReAgg framework. **(a)** Shared feature extraction from frozen reference models. **(b)** Token reliability weighting combining relative position and hidden discrepancy ($w_i = p_i \eta_i$). **(c)** Carrier-specific aggregation producing the reweighted document score $S^{(D)}_{\text{cal}}$.

carrier D, token alignment, context truncation, and validity masks determine the valid position set $\mathcal{I}_D \subseteq \{1, \ldots, T\}$ with effective length $T_D = |\mathcal{I}_D|$. The carrier then extracts token-level evidence:

$$\mathbf{z}^{(D)}_{\mathcal{I}_D} = f_D(x) = \left(z_i^{(D)}\right)_{i \in \mathcal{I}_D}. \tag{1}$$

Here, f_D is the carrier-specific evidence extraction function, and $z_i^{(D)}$ denotes the local diagnostic evidence at position i (which may consist of scalar statistics or paired components for ratio-based scores, as specified in Appendix D).

We write the raw carrier score by passing unit weights to its aggregation operator:

$$S^{(D)}_{\text{raw}} = \mathcal{A}_D\left(\mathbf{z}^{(D)}_{\mathcal{I}_D}, \mathbf{1}_{\mathcal{I}_D}\right). \tag{2}$$

Here, $\mathcal{A}_D$ is the carrier-specific aggregation operator, and $\mathbf{1}_{\mathcal{I}_D}$ is the all-ones vector of length T_D. ReAgg replaces the unit weights with non-negative reliability weights:

$$S^{(D)}_{\text{cal}} = \mathcal{A}_D\left(\mathbf{z}^{(D)}_{\mathcal{I}_D}, \mathbf{w}_{\mathcal{I}_D}\right), \tag{3}$$

where $\mathbf{w}_{\mathcal{I}_D} = (w_i)_{i \in \mathcal{I}_D}$ contains reliability weights satisfying $w_i \geq 0$ and $\sum_{i \in \mathcal{I}_D} w_i > 0$, and $S^{(D)}_{\text{cal}}$ is the reweighted document score. Equations (2) and (3) separate local evidence generation from document-level contribution assignment. As illustrated in Figure 2, shared feature extraction (a) produces carrier evidence and aligns representations at valid positions; token reweighting (b) constructs reliability weights from observable cues; and document scoring (c) aggregates the unchanged evidence using these weights without additional model training.

3.2 Variance Analysis of Evidence Aggregation

We fix carrier D and effective length $n = T_D$, indexing valid positions as $1, \ldots, n$. Let $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ denote random local evidence and C denote observable conditioning information used to construct the weights. For class label $Y \in \{0, 1\}$, consider the conditional mean model:

$$Z_i = \mu_Y(C) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid Y, C] = 0, \tag{4}$$

where $\mu_Y(C)$ is shared across positions, and ε_i has finite conditional second moment. This working model isolates the effect of weighting on within-class aggregation variance.

Define normalized weights $a_i = w_i / \sum_j w_j$ and $\mathbf{a} = (a_1, \ldots, a_n)^\top$ satisfying $\mathbf{1}^\top \mathbf{a} = 1$. We assume that $\mathbf{a}$ is measurable with respect to C, so the weights are fixed under conditioning. The linear score $S_\mathbf{a} = \mathbf{a}^\top \mathbf{Z}$ satisfies $\mathbb{E}[S_\mathbf{a} \mid Y, C] = \mu_Y(C)$, ensuring conditional unbiasedness. Its pooled within-class variance is $V(\mathbf{a}) = \mathbf{a}^\top \bar{\Sigma} \mathbf{a}$, where $\bar{\Sigma} = \sum_{y=0}^{1} q_y \operatorname{Cov}(\mathbf{Z} \mid Y = y, C)$ for fixed class weights $q_y \geq 0$ with $\sum_y q_y = 1$. We omit conditioning on C below for brevity. Under the diagonal working model $\bar{\Sigma} = \operatorname{diag}(v_1, \ldots, v_n)$ with position variances $v_i > 0$, generalized least squares (GLS) yields the non-negative oracle precision reference (Appendix A.2):

$$\pi_i^\star = v_i^{-1}, \qquad a_i^\star = \frac{\pi_i^\star}{\sum_{j=1}^n \pi_j^\star}, \qquad V^\star = \frac{1}{\sum_{j=1}^n \pi_j^\star}. \tag{5}$$

Here, $\pi_i^\star$ is the ideal precision, $a_i^\star$ is the oracle weight at position i, and $V^\star$ is the minimum achievable variance.

Proposition 1 (Exact Variance Excess Identity). *Under the diagonal working model, any non-negative weight vector* $\mathbf{a}$ *satisfying* $\mathbf{1}^\top \mathbf{a} = 1$ *obeys*

$$V(\mathbf{a}) - V^\star = \sum_{i=1}^{n} v_i (a_i - a_i^\star)^2 = V^\star \chi^2(\mathbf{a} \| \mathbf{a}^\star), \tag{6}$$

where $\chi^2(\mathbf{a} \| \mathbf{a}^\star) = \sum_{i=1}^n (a_i - a_i^\star)^2 / a_i^\star$ *is the Pearson* χ^2*-divergence.*

Proposition 1 implies that a weight vector $\mathbf{a}$ has lower variance than uniform weighting $\mathbf{a}_{\text{unif}} = \mathbf{1}/n$ if and only if $\chi^2(\mathbf{a} \| \mathbf{a}^\star) < \chi^2(\mathbf{a}_{\text{unif}} \| \mathbf{a}^\star)$. Equation (6) establishes the exact variance cost of departing from the oracle allocation. The proof is provided in Appendix A.3.

Zero Weights and Support Loss. When some positions receive zero weight, let $A = \{i : w_i > 0\}$ denote the positive support and let $M_A = \sum_{i \in A} a_i^\star \leq 1$ denote the retained oracle weight mass. Decomposing the variance gives (Appendix A.4):

$$\frac{V(\mathbf{a})}{V^\star} = \frac{1}{M_A} \left[1 + \chi^2(\mathbf{a}_A \| \mathbf{a}_A^\star) \right], \tag{7}$$

where $\mathbf{a}_A = (a_i)_{i \in A}$ and $\mathbf{a}_A^\star = (a_i^\star / M_A)_{i \in A}$. Equation (7) separates the variance cost of support restriction ($1/M_A \geq 1$) from the weighting mismatch within the retained support. A log-error bound for proxy weights is provided in Appendix A.5.

3.3 Computable Reliability Surrogates and Weight Construction

The oracle precisions in Section 3.2 are unavailable from a single text. ReAgg instead derives weights from relative position and hidden discrepancy (Figure 2(b)).

Relative Position Prior. Let $\kappa(i)$ denote the order of position i within $\mathcal{I}_D$. We define

$$u_i = \frac{\kappa(i)}{T_D}, \qquad p_i = u_i^\beta, \qquad \beta \geq 0. \tag{8}$$

The position factor emphasizes later positions when $\beta > 0$ and becomes uniform when $\beta = 0$.

Hidden-State Discrepancy. Let $h_{i,L}^{\text{base}}$ and $h_{i,L}^{\text{inst}}$ denote the aligned hidden states from the final layer L of the frozen base and instruction-tuned models at the prefix position predicting token i. We define

$$d_i = 1 - \cos\left(h_{i,L}^{\text{base}}, h_{i,L}^{\text{inst}} \right). \tag{9}$$

We normalize discrepancies by ascending rank within the valid positions:

$$r_i = \operatorname{RankNorm}_{\mathcal{I}_D}(d)_i \in [0, 1]. \tag{10}$$

The low-discrepancy factor favors positions with smaller representational changes; the high-discrepancy factor provides a directional control:

$$\eta_i^{\text{low}} = (1 - r_i)^\gamma, \qquad \eta_i^{\text{high}} = r_i^\gamma, \qquad \gamma \geq 0. \tag{11}$$

Rank normalization, ties, and boundary conventions are specified in Appendix D.3.

Joint Weights. The position and low-discrepancy factors combine multiplicatively:

$$w_i = p_i \eta_i^{\text{low}} = u_i^{\beta}(1 - r_i)^{\gamma}. \tag{12}$$

ReAgg Fixed uses $\beta = \gamma = 1$ and the low-discrepancy branch, abbreviated as $\eta_i = \eta_i^{\text{low}}$ in Figure 2. These observable cues provide a computable weighting rule motivated by the variance analysis, without estimating the oracle precisions.

3.4 Carrier-Specific Aggregation Operators

Following Figure 2(c), ReAgg passes unchanged local evidence and the joint weights to the carrier's aggregation operator. Let $W_D = \sum_{i \in \mathcal{I}_D} w_i > 0$.

Weighted Mean. For Likelihood, Entropy, Log-Rank, and IRM-mean,

$$S_{\text{mean}}^{(D)} = \frac{1}{W_D} \sum_{i \in \mathcal{I}_D} w_i z_i^{(D)}. \tag{13}$$

This is the normalized linear aggregation analyzed in Section 3.2 under its stated assumptions.

Shared-Weight Ratio. For LRR and DNA-style ratio, the same weights apply to the local numerator and denominator components:

$$S_{\text{ratio}}^{(D)} = \frac{\sum_{i \in \mathcal{I}_D} w_i n_i^{(D)}}{c_D \sum_{i \in \mathcal{I}_D} w_i m_i^{(D)}}. \tag{14}$$

Here, $n_i^{(D)}$ and $m_i^{(D)}$ are the local numerator and denominator components, with $c_D = 1$ for LRR and $c_D = 2$ for DNA-style ratio. Both components use the same valid positions, and the weighted denominator must be nonzero.

Unnormalized Weighted Sum. For IRM-sum, we aggregate without weight normalization:

$$S_{\text{sum}}^{(D)} = \sum_{i \in \mathcal{I}_D} w_i z_i^{(D)} = W_D S_{\text{mean}}^{(D)}. \tag{15}$$

The score depends on both relative token contributions and the total weight W_D.

Weighted Spectral Statistic. For Fourier-style detection, let q_i be token surprisal standardized within the valid sequence before weighting. Preserving token order, we center and weight the sequence before measuring its low-frequency energy:

$$\widetilde{q}_i = \frac{w_i}{\bar{w}}(q_i - \mu_w), \qquad S_{\text{Fourier}}^{(D)} = \frac{1}{T_D} \sum_{k=1}^{K} \left| \text{rFFT}(\widetilde{\mathbf{q}})_k \right|^2. \tag{16}$$

Here, $\mu_w = \sum_{i \in \mathcal{I}_D} w_i q_i / W_D$, $\bar{w} = W_D / T_D$, and $K = \min(10, \lfloor T_D/2 \rfloor)$. The real-input FFT uses an unnormalized forward transform; the sum excludes the DC coefficient.

Operator properties and carrier evidence definitions appear in Appendices C and D.

4 Experiments

We evaluate ReAgg through three research questions: **RQ1:** Does fixed weighting improve detection across datasets and carriers? **RQ2:** Do position and hidden-state discrepancy provide complementary weighting cues? **RQ3:** When does ReAgg help or fail?

4.1 Experimental Setup

Data and Reference Models. We evaluate ReAgg on 18 subsets from five benchmarks: DetectRL (Wu et al., 2024), M4 (Wang et al., 2024), EvoBench (Yu et al., 2025), MAGE (Li et al., 2024), and RealDet (Zhu et al., 2025), comprising 24,210 text records. The primary comparison covers 13 subsets, while the remaining five provide supplementary analyses. We use frozen Falcon-7B and Falcon-7B-Instruct reference models to extract aligned token outputs and last-layer hidden states. Subset composition, evaluation splits, and limitations of source metadata are detailed in Appendix E.

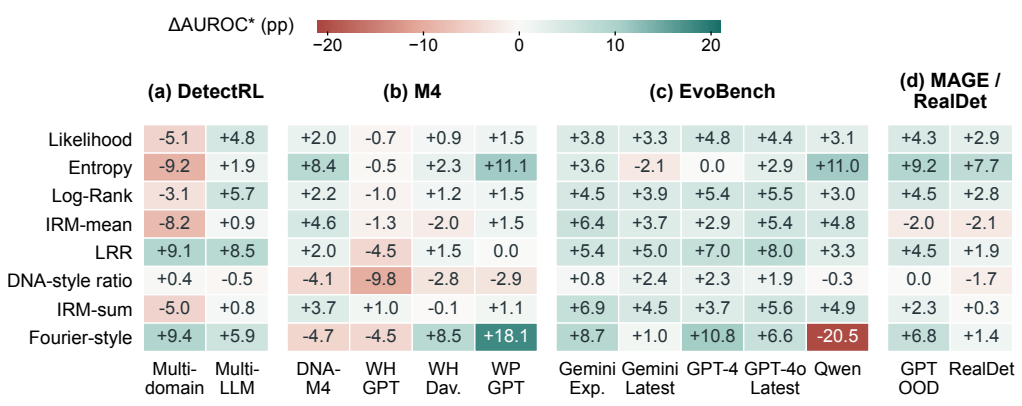


Figure 3: ReAgg Fixed gains over Raw in direction-corrected AUROC (percentage points) across the 13 primary subsets, averaged over five splits. In panel (b), WH, WP, GPT, and Dav. denote WikiHow, Wikipedia, ChatGPT, and Davinci, respectively.

Comparison Settings. We compare the original aggregation (**Raw**) against **ReAgg Fixed** across the eight carriers introduced in Section 3.4, spanning mean, ratio, sum, and spectral structures. ReAgg Fixed uses one shared configuration ($\beta = \gamma = 1$, low-discrepancy branch) across all subsets and carriers without per-subset configuration selection. Within each carrier, both settings share reference models, local evidence, valid positions, and data splits; only the aggregation weights change. ReAgg requires no additional model training. Carrier implementations are detailed in Appendix D, and configurations and development context in Appendix E.

Evaluation Metrics. We report direction-corrected AUROC, $\mathrm{AUROC}^* = \max(\mathrm{AUROC}, 1 - \mathrm{AUROC})$. For each setting and evaluation split, evaluation labels determine the score direction. This metric measures ranking separability rather than performance with a direction fixed before testing. For the primary comparison, we first average over five random splits within each subset–carrier combination, then compute an equally weighted macro-average across the 13 primary subsets and eight carriers, reporting absolute gains over Raw and positive combination counts. A supplementary evaluation selects each score direction on development data and holds it fixed during testing; results and paired uncertainty analysis are reported in Appendix E.11.

4.2 RQ1: Detection Performance across Datasets and Carriers

Table 1 reports direction-corrected AUROC across the five benchmark families and overall, while Figure 3 illustrates the corresponding subset-level gains.

Overall Performance. ReAgg Fixed increases macro-average direction-corrected AUROC from 75.22% to 77.54% (+2.32 percentage points), improving 79 of the 104 subset–carrier combinations with positive overall gains for seven of the eight carriers. In the supplementary evaluation selecting directions on development data and holding them fixed during testing, test AUROC similarly increases from 74.15% to 77.02% (+2.87 percentage points, Appendix E.11).

Performance Patterns and Variations. Average performance improves across all five benchmark families. While gains vary across detector carriers—with DNA-style ratio being the sole carrier-level exception (-1.10 percentage points)—positive overall gains occur across carriers in all four scoring structures, supporting the use of shared weighting with carrier-specific aggregation.

4.3 RQ2: Contributions of Position and Hidden-State Discrepancy

Table 2 evaluates the weighting design through single-cue ablations, directional controls, and within-document permutations. All comparisons follow the primary protocol in Section 4.1 with unit exponents and five permutation seeds.

Table 1: Direction-corrected AUROC (%) for Raw and ReAgg Fixed. Family columns average their constituent subsets; Overall averages all 13 primary subsets. Macro-average additionally averages across carriers. Bold marks the higher Overall score in each pair. Gains use unrounded results.

Carrier	Setting	DetectRL	M4	EvoBench	MAGE	RealDet	Overall
Likelihood	Raw	59.89	89.43	68.72	78.10	85.01	75.71
	+ReAgg (ours)	59.74	90.34	72.59	82.39	87.95	**78.01**
Entropy	Raw	64.05	80.72	55.85	61.79	75.44	66.73
	+ReAgg (ours)	60.38	86.05	58.94	70.97	83.18	**70.29**
Log-Rank	Raw	59.42	90.08	68.50	78.36	85.98	75.85
	+ReAgg (ours)	60.72	91.03	72.97	82.90	88.76	**78.62**
IRM-mean	Raw	79.59	89.06	74.50	92.63	85.46	82.00
	+ReAgg (ours)	75.91	89.77	79.13	90.63	83.31	**83.11**
LRR	Raw	58.65	89.40	65.09	76.82	87.14	74.18
	+ReAgg (ours)	67.43	89.16	70.84	81.33	89.05	**78.16**
DNA-style ratio	Raw	88.03	89.98	84.65	95.03	94.17	**88.34**
	+ReAgg (ours)	87.95	85.07	86.08	95.04	92.52	87.24
IRM-sum	Raw	87.23	81.57	70.39	80.50	77.73	77.76
	+ReAgg (ours)	85.11	82.99	75.51	82.83	77.98	**80.04**
Fourier-style	Raw	57.87	60.06	61.31	61.97	70.64	61.16
	+ReAgg (ours)	65.53	64.40	62.64	68.81	72.01	**64.82**
Macro-average	Raw	69.34	83.79	68.63	78.15	82.70	75.22
	+ReAgg (ours)	70.35	84.85	72.34	81.86	84.35	**77.54**

Table 2: Weighting ablations and controls. Scores are macro-average direction-corrected AUROC (%); Δ is the gain over Raw in percentage points. Here, $p_i = u_i$, $\eta_i = 1 - r_i$, and π denotes a within-document permutation.

Weighting setting	Token weight	AUROC*	Δ (pp)
Raw	1	75.22	—
Position-only	p_i	76.09	$+0.87$
Hidden-low-only	η_i	76.19	$+0.97$
ReAgg (ours)	$p_i\eta_i$	**77.54**	$\mathbf{+2.32}$
Joint-high	$p_i r_i$	73.17	-2.05
Joint-weight shuffle	$(p\eta)_{\pi(i)}$	73.28	-1.93
Position shuffle	$p_{\pi(i)}\eta_i$	75.33	$+0.11$
Hidden shuffle	$p_i\eta_{\pi(i)}$	75.22	≈ 0

Cue Complementarity and Discrepancy Direction. Joint weighting reaches 77.54% macro-average AUROC*, exceeding Position-only and Hidden-low-only by 1.45 and 1.35 percentage points, confirming cue complementarity. Replacing the low-discrepancy branch with the high-discrepancy branch reduces the score to 73.17%; the low branch performs better in 90 of 104 combinations, validating our default direction.

Weight Assignment to Tokens. Joint-weight shuffling preserves weight values and concentration while permuting positions; macro-average AUROC* drops to 73.28% (4.26 percentage points below ReAgg), supporting the importance of weight-to-token assignment beyond non-uniformity alone. Shuffling position or hidden factors individually also degrades performance (75.33% and 75.22%), though these permutations can additionally alter total product weight and affect IRM-sum's scale.

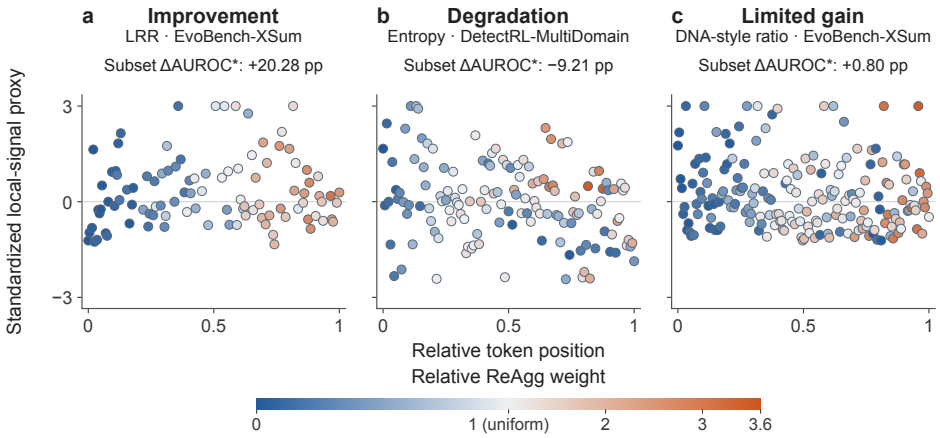


Figure 4: Token-level weight allocations in three illustrative cases: **(a)** marked improvement in LRR; **(b)** degradation in Entropy; and **(c)** limited gain in DNA-style ratio. Points show standardized local-signal proxies clipped to $[-3, 3]$; colors indicate relative ReAgg weights, with 1 denoting uniform weighting. Undefined proxy values are omitted. Annotated $\Delta\mathrm{AUROC}^*$ gains are subset-level results; EvoBench-XSum belongs to the supplementary scope. Proxy definitions and visualization details are in Appendix F.

4.4 RQ3: Under What Conditions Does ReAgg Gain or Degrade?

Figure 4 illustrates token-level weight allocations across three representative cases; supplementary subset EvoBench-XSum is excluded from primary averages.

Gains across Carriers. On EvoBench-XSum, LRR's direction-corrected AUROC increases from 52.10% to 72.38%, a gain of 20.28 percentage points; DNA-style ratio further improves from an already high baseline of 96.69% to 97.48% (+0.80 points). These cases illustrate that the same weighting rule provides gains across different baseline separability levels and carrier structures.

Variations and Failure. On DetectRL-MultiDomain, Entropy's direction-corrected AUROC decreases from 74.11% to 64.90%, demonstrating that the efficacy of fixed cues depends on carrier and data conditions. Figure 4 juxtaposes these outcomes with specific weight allocations, providing concrete case references for understanding heterogeneous performance; local-signal proxies in the figure are not direct measurements of conditional variances. Beyond these illustrative cases, Table 1 shows that LRR and Fourier-style achieve the largest overall gains among the evaluated carriers, making them promising choices for ReAgg within the studied settings.

5 Conclusion

We introduced ReAgg, a training-free evidence aggregation framework for machine-generated text detection. While preserving local evidence, ReAgg constructs reliability weights from relative position and base–instruction representation discrepancies across diverse scoring structures. Under a diagonal covariance model, an exact variance excess identity characterizes the cost of weighting mismatch and support loss. Across 13 primary subsets and eight carriers, ReAgg improves macro-average direction-corrected AUROC from 75.22% to 77.54%; with directions selected on development data, macro-average test AUROC increases from 74.15% to 77.02%. Ablations and controls confirm cue complementarity and weight-to-token alignment, establishing evidence aggregation as a principled design dimension for statistical text detection.

Reproducibility Statement

This study uses public benchmark datasets and frozen open-weight reference models. The main evaluation applies a unified ReAgg configuration across all subsets and carriers without additional model training; Raw and ReAgg comparisons within each carrier share reference models, local evidence, valid positions, and data splits, varying only the aggregation weights. Theoretical derivations and operator analyses are provided in Appendices A–C, implementation details in Appendix D, data splits and supplementary evaluations in Appendix E, and visualization definitions in Appendix F. Anonymized code will be submitted with supplementary materials to support method reproduction and result verification.

AI Use Statement

This research was led by the authors in problem formulation, method design, mathematical derivation, experimental implementation, and result analysis, with derivation processes, experimental code, and experiment logs serving as the basis for the reported conclusions. Generative AI tools assisted in these efforts and in manuscript preparation, including providing feedback on theoretical derivations and experimental methodology, assisting with result interpretation, and supporting literature search, manuscript drafting and editing, LaTeX formatting, and figure preparation. The authors reviewed AI suggestions, decided on their adoption based on original references, mathematical derivations, and experimental records, and verified the technical correctness, citation accuracy, and consistency of the final content with the experimental evidence. Research judgment and final decisions were made by the authors, who bear full responsibility for the paper's content.

Ethics Statement

This study uses publicly available benchmark datasets and open-weight reference models for machine-generated text detection research, involving no new human participant recruitment or data collection. The work aims to provide technical support for text provenance analysis. Detection performance varies across data conditions and carriers; detection scores should therefore serve as auxiliary evidence rather than definitive determinations of text provenance. In practice, detection results should be combined with provenance records, usage context, and human review to avoid making adverse decisions (such as academic misconduct determinations) based solely on detection outputs. Use and sharing of data, models, and supplementary materials should follow the corresponding licenses and access conditions.

References

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations (ICLR)*, 2024.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12463–12492, 2024. doi: 10.18653/v1/2024.acl-long.674.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. In *ACL System Demonstrations*, 2019.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, et al. Spotting llms with binoculars: Zero-shot detection of machine-generated text. In *International Conference on Machine Learning (ICML)*, 2024.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Lei Jiang, Desheng Wu, Xiaolong Zheng, et al. Leveraging human and machine preferences for zero-shot detection of ai-generated text. In *Findings of ACL*, 2026.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36–53. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.3. URL https://aclanthology.org/2024.acl-long.3/.

Runheng Liu, Heyan Huang, Xingchen Xiao, and Zhijing Wu. Zero-shot detection of LLM-generated text via implicit reward model. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f50258b34f1c5080e43281e05050034e-Abstract-Conference.html.

Haitong Luo, Weiyao Zhang, Suhang Wang, Wenji Zou, Chungang Lin, Xuying Meng, and Yujun Zhang. SpecDetect: Simple, fast, and training-free detection of LLM-generated text via spectral analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):32356–32364, 2026. doi: 10.1609/aaai.v40i38.40510. URL https://ojs.aaai.org/index.php/AAAI/article/view/40510.

Shixuan Ma and Quan Wang. Zero-shot detection of llm-generated text using token cohesiveness. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Pablo Miralles-González, Javier Huertas-Tato, Alejandro Martín, et al. Not all tokens are created equal: Perplexity attention weighted networks for ai generated text detection. *arXiv preprint arXiv:2501.03940*, 2025.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, et al. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning (ICML)*, 2023.

Irene Solaiman, Jack Clark, Miles Brundage, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of EMNLP*, 2023.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, et al. Ghostbuster: Detecting text ghostwritten by large language models. *arXiv preprint arXiv:2305.15047*, 2023.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 1369–1407, 2024.

Junchao Wu, Runzhe Zhan, Derek F. Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S. Chao. DetectRL: Benchmarking LLM-generated text detection in real-world scenarios. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Junxi Wu, Kailin Huang, Dongjian Hu, Bin Chen, Hao Wu, Shu-Tao Xia, and Changliang Zou. Alignment imprint: Zero-shot AI-generated text detection via provable preference discrepancy, 2026. URL https://arxiv.org/abs/2604.16923.

Yihuai Xu, Yongwei Wang, Yifei Bi, et al. Training-free llm-generated text detection by mining token probability sequences. In *International Conference on Learning Representations (ICLR)*, 2025.

Sungduk Yu, Man Luo, Avinash Madasu, Vasudev Lal, and Phillip Howard. Is your paper being reviewed by an LLM? Benchmarking AI text detection in peer review. In *International Conference on Learning Representations (ICLR)*, 2026.

Xiao Yu, Yi Yu, Dongrui Liu, Kejiang Chen, Weiming Zhang, Nenghai Yu, and Jing Shao. EvoBench: Towards real-world LLM-generated text detection benchmarking for evolving large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14605–14620, 2025.

Xiaowei Zhu, Yubing Ren, Yanan Cao, Xixun Lin, Fang Fang, and Yangxi Li. Reliably bounding false positives: A zero-shot machine-generated text detection framework via multiscaled conformal prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 12298–12319, 2025.

Xiaowei Zhu, Yubing Ren, Fang Fang, et al. Exons-detect: Identifying and amplifying exonic tokens via hidden-state discrepancy for robust ai-generated text detection. In *Association for Computational Linguistics (ACL)*, 2026.

A Theoretical Derivations and Proofs

A.1 Conditional Model and Variance Criterion

We fix a carrier D and effective length $n = T_D$. Let $\mathbf{Z} = (Z_1, \dots, Z_n)^\top$ denote the random vector of local evidence. Under the working model in Eq. (4),

$$Z_i = \mu_Y(C) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid Y, C] = 0, \qquad \mathbb{E}[\varepsilon_i^2 \mid Y, C] < \infty. \tag{A.1}$$

Here, C contains observable weighting conditions, and the conditional mean $\mu_Y(C)$ is shared across valid positions. The weights $\mathbf{a}$ are measurable with respect to C and satisfy $a_i \geq 0$ and $\mathbf{1}^\top \mathbf{a} = 1$.

Since the weights are fixed under conditioning,

$$\mathbb{E}[S_{\mathbf{a}} \mid Y, C] = \sum_i a_i \mathbb{E}[Z_i \mid Y, C] = \mu_Y(C), \tag{A.2}$$

$$\mathrm{Var}(S_{\mathbf{a}} \mid Y = y, C) = \mathbf{a}^\top \Sigma_y(C) \mathbf{a}, \tag{A.3}$$

where $S_{\mathbf{a}} = \mathbf{a}^\top \mathbf{Z}$ and $\Sigma_y(C) = \mathrm{Cov}(\mathbf{Z} \mid Y = y, C)$.

For fixed class weights $q_y \geq 0$ with $\sum_y q_y = 1$, define

$$\bar{\Sigma}(C) = \sum_y q_y \Sigma_y(C). \tag{A.4}$$

The pooled within-class variance is then

$$V(\mathbf{a} \mid C) = \sum_y q_y \,\mathrm{Var}(S_{\mathbf{a}} \mid Y = y, C) = \mathbf{a}^\top \bar{\Sigma}(C) \mathbf{a}. \tag{A.5}$$

The common conditional mean isolates the variance effect of reweighting. With position-specific conditional means, weighting can also change the score mean. Conditioning is on observable weighting information, rather than the complete input text, which would determine the evidence under frozen models. Diagonal covariance requires zero pooled off-diagonal covariances, without requiring independence of all token variables. The results below concern normalized linear aggregation under these assumptions; they do not by themselves establish an AUROC improvement.

A.2 Full-Covariance GLS and Non-Negative Optimization

We suppress conditioning on C and assume $\bar{\Sigma} \succ 0$. First consider the equality-constrained problem

$$\min_{\mathbf{a} \in \mathbb{R}^n} \mathbf{a}^\top \bar{\Sigma} \mathbf{a} \qquad \text{subject to} \quad \mathbf{1}^\top \mathbf{a} = 1. \tag{A.6}$$

Its Lagrangian is

$$\mathcal{L}(\mathbf{a}, \lambda) = \mathbf{a}^\top \bar{\Sigma} \mathbf{a} - \lambda(\mathbf{1}^\top \mathbf{a} - 1). \tag{A.7}$$

Stationarity gives $2\bar{\Sigma}\mathbf{a} = \lambda \mathbf{1}$. Applying the normalization constraint yields

$$\mathbf{a}^\star_{\mathrm{GLS}} = \frac{\bar{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^\top \bar{\Sigma}^{-1} \mathbf{1}}, \qquad V^\star_{\mathrm{GLS}} = \frac{1}{\mathbf{1}^\top \bar{\Sigma}^{-1} \mathbf{1}}. \tag{A.8}$$

For any normalized $\mathbf{a}$, let $\mathbf{b} = \mathbf{a} - \mathbf{a}^\star_{\mathrm{GLS}}$. Then $\mathbf{1}^\top \mathbf{b} = 0$ and $\bar{\Sigma}\mathbf{a}^\star_{\mathrm{GLS}} = V^\star_{\mathrm{GLS}}\mathbf{1}$. Expanding the quadratic form gives

$$V(\mathbf{a}) - V^\star_{\mathrm{GLS}} = (\mathbf{a} - \mathbf{a}^\star_{\mathrm{GLS}})^\top \bar{\Sigma}(\mathbf{a} - \mathbf{a}^\star_{\mathrm{GLS}}). \tag{A.9}$$

The GLS solution may contain negative entries. With non-negativity, the problem becomes

$$\min_{\mathbf{a}\geq\mathbf{0},\, \mathbf{1}^\top\mathbf{a}=1} \mathbf{a}^\top \bar{\Sigma}\mathbf{a}. \tag{A.10}$$

Strict convexity gives a unique minimizer $\mathbf{a}^\star_+$. The Karush–Kuhn–Tucker conditions are

$$\begin{aligned} 2\bar{\Sigma}\mathbf{a}^\star_+ &= \lambda\mathbf{1} + \boldsymbol{\nu}, \\ \mathbf{a}^\star_+ \geq \mathbf{0}, &\qquad \mathbf{1}^\top\mathbf{a}^\star_+ = 1, \\ \boldsymbol{\nu} \geq \mathbf{0}, &\qquad \nu_i(a^\star_+)_i = 0. \end{aligned} \tag{A.11}$$

If $\mathbf{a}^\star_{\mathrm{GLS}} \geq \mathbf{0}$, it also solves this problem. Otherwise, the non-negative optimum lies on a boundary of the simplex.

For any feasible $\mathbf{a}$, expansion around $\mathbf{a}^\star_+$ gives

$$V(\mathbf{a}) - V(\mathbf{a}^\star_+) = (\mathbf{a} - \mathbf{a}^\star_+)^\top \bar{\Sigma}(\mathbf{a} - \mathbf{a}^\star_+) + \boldsymbol{\nu}^\top\mathbf{a}. \tag{A.12}$$

The linear term follows from $\mathbf{1}^\top(\mathbf{a} - \mathbf{a}^\star_+) = 0$ and complementary slackness. Both terms are non-negative.

These full-covariance solutions require a covariance model. The diagonal model below supplies an analytical reference for ReAgg's observable weighting rule, rather than a covariance estimate computed by the method.

A.3 Diagonal Oracle and Proof of Proposition 1

Let $\bar{\Sigma} = \mathrm{diag}(v_1, \ldots, v_n)$ with $v_i > 0$. Define

$$\pi^\star_i = v_i^{-1}, \qquad P = \sum_i \pi^\star_i, \qquad a^\star_i = \frac{\pi^\star_i}{P}, \qquad V^\star = \frac{1}{P}. \tag{A.13}$$

The full-covariance solution reduces to these weights under the diagonal model. They are strictly positive and therefore solve both the equality-constrained and non-negative problems. They satisfy

$$v_i a^\star_i = V^\star, \qquad \bar{\Sigma}\mathbf{a}^\star = V^\star\mathbf{1}. \tag{A.14}$$

For any normalized non-negative $\mathbf{a}$, set $\mathbf{b} = \mathbf{a} - \mathbf{a}^\star$. Since $\mathbf{1}^\top\mathbf{b} = 0$,

$$\begin{aligned} V(\mathbf{a}) &= (\mathbf{a}^\star + \mathbf{b})^\top \bar{\Sigma}(\mathbf{a}^\star + \mathbf{b}) \\ &= V^\star + 2V^\star\mathbf{1}^\top\mathbf{b} + \mathbf{b}^\top\bar{\Sigma}\mathbf{b} \\ &= V^\star + \sum_i v_i(a_i - a^\star_i)^2. \end{aligned} \tag{A.15}$$

Substituting $v_i = V^\star/a^\star_i$ proves

$$V(\mathbf{a}) - V^\star = V^\star \sum_i \frac{(a_i - a^\star_i)^2}{a^\star_i} = V^\star\chi^2(\mathbf{a}\|\mathbf{a}^\star). \tag{A.16}$$

This establishes Eq. (6) and proves Proposition 1 in the main text.

Because $V^\star > 0$, comparing two normalized weight vectors is equivalent to comparing their divergences from the same oracle. In particular,

$$V(\mathbf{a}) < V(\mathbf{a}_{\mathrm{unif}}) \quad\Longleftrightarrow\quad \chi^2(\mathbf{a}\|\mathbf{a}^\star) < \chi^2(\mathbf{a}_{\mathrm{unif}}\|\mathbf{a}^\star), \tag{A.17}$$

where $\mathbf{a}_{\mathrm{unif}} = \mathbf{1}/n$. Thus, non-uniform weighting helps under this criterion precisely when it is closer to the oracle in Pearson divergence.

A.4 Proof of the Support-Set Decomposition

Let $A = \{i : w_i > 0\}$ be nonempty and $a_i = w_i / \sum_j w_j$. Define

$$M_A = \sum_{i \in A} a_i^\star, \qquad \mathbf{a}_A = (a_i)_{i \in A}, \qquad \mathbf{a}_A^\star = \left(\frac{a_i^\star}{M_A} \right)_{i \in A}. \tag{A.18}$$

Since A is nonempty and every oracle weight is positive, $0 < M_A \leq 1$. Both restricted vectors sum to one, and $a_i = 0$ outside A.

Using $v_i = V^\star / a_i^\star$,

$$\begin{aligned} \frac{V(\mathbf{a})}{V^\star} &= \sum_{i \in A} \frac{a_i^2}{a_i^\star} \\ &= \frac{1}{M_A} \sum_{i \in A} \frac{a_i^2}{a_{A,i}^\star} \\ &= \frac{1}{M_A} \left[1 + \chi^2(\mathbf{a}_A \| \mathbf{a}_A^\star) \right]. \end{aligned} \tag{A.19}$$

The last step uses $\sum_i a_i^2 / b_i = 1 + \chi^2(\mathbf{a} \| \mathbf{b})$ for probability vectors with $b_i > 0$. This proves Eq. (7) in the main text.

Equivalently, the excess variance separates as

$$V(\mathbf{a}) - V^\star = V^\star \frac{1 - M_A}{M_A} + \frac{V^\star}{M_A} \chi^2(\mathbf{a}_A \| \mathbf{a}_A^\star). \tag{A.20}$$

The first term is the cost of discarding oracle weight mass. The second measures relative weighting mismatch among retained positions. The minimum variance attainable on A is $V^\star / M_A$, achieved by $\mathbf{a}_A^\star$. This cost depends on the oracle mass retained, rather than only on the number of retained positions. It compares support restriction with the full-support oracle; it does not imply that every full-support weighting has lower variance than every sparse weighting.

A.5 Log-Error Bound for Proxy Weights

Under the same diagonal model, consider positive unnormalized weights on A. Suppose there exist $c_0 > 0$ and $\delta \geq 0$ such that

$$\left| \log \frac{w_i}{c_0 \pi_i^\star} \right| \leq \delta, \qquad i \in A. \tag{A.21}$$

Write $w_i = c_0 \pi_i^\star t_i$, where $e^{-\delta} \leq t_i \leq e^\delta$. Let $b_i = a_{A,i}^\star$ and let $\mathbb{E}_b$ denote expectation under the probability masses $(b_i)_{i \in A}$. Normalization gives

$$a_i = \frac{b_i t_i}{\mathbb{E}_b[t]}, \qquad \frac{V(\mathbf{a})}{V^\star} = \frac{1}{M_A} \frac{\mathbb{E}_b[t^2]}{\mathbb{E}_b[t]^2}. \tag{A.22}$$

We prove the required moment inequality directly. For any positive random variable $t \in [m, M]$, with $0 < m \leq M$, the inequality $(t - m)(M - t) \geq 0$ implies

$$\mathbb{E}[t^2] \leq (m + M)\mu - mM, \qquad \mu = \mathbb{E}[t]. \tag{A.23}$$

Furthermore,

$$\frac{(m+M)^2}{4mM} - \frac{(m+M)\mu - mM}{\mu^2} = \frac{\big((m+M)\mu - 2mM\big)^2}{4mM\mu^2} \geq 0. \tag{A.24}$$

Together with $\mathbb{E}[t^2] \geq \mu^2$, this yields the Kantorovich-type bound

$$1 \leq \frac{\mathbb{E}[t^2]}{\mathbb{E}[t]^2} \leq \frac{(m+M)^2}{4mM}. \tag{A.25}$$

Setting $m = e^{-\delta}$ and $M = e^{\delta}$ gives

$$\frac{(m+M)^2}{4mM} = \left(\frac{e^{\delta} + e^{-\delta}}{2}\right)^2 = \cosh^2(\delta). \tag{A.26}$$

Substitution into Eq. (A.22) proves

$$\frac{1}{M_A} \leq \frac{V(\mathbf{a})}{V^\star} \leq \frac{\cosh^2(\delta)}{M_A}. \tag{A.27}$$

For full support, $M_A = 1$. For $\delta = 0$, the weights are proportional to the oracle precision on A, so the lower and upper bounds coincide at $1/M_A$. For $\delta > 0$, the moment bound is sharp over distributions supported on $\{m, M\}$: equality is attained when the mass at m is $M/(m+M)$ and the mass at M is $m/(m+M)$, giving $\mathbb{E}[t] = 2mM/(m+M)$. A fixed finite support need not admit these exact masses. The bound is conditional on the stated uniform log-error assumption; it does not estimate the proxy error of a particular weighting rule.

A.6 Log-Space Parameterization and an Oracle Prediction Target

To formalize a precision surrogate, fix carrier D and length n, draw a document from a population distribution, and sample a valid position J uniformly. Let

$$\Lambda = \log \pi^\star_J, \qquad X = (U, R) = (u_J, r_J). \tag{A.28}$$

Here, Λ is the unobserved log-precision, while X contains the two observable weighting cues. Assume $\mathbb{E}[\Lambda^2] < \infty$.

Among square-integrable functions of X, the conditional expectation

$$g^\star(X) = \mathbb{E}[\Lambda \mid X] \tag{A.29}$$

minimizes squared prediction error. For any square-integrable $g(X)$,

$$\begin{aligned}\mathbb{E}\big[(\Lambda - g(X))^2\big] = \mathbb{E}\big[(\Lambda - g^\star(X))^2\big] \\ + \mathbb{E}\big[(g^\star(X) - g(X))^2\big].\end{aligned} \tag{A.30}$$

To see this, expand $\Lambda - g = (\Lambda - g^\star) + (g^\star - g)$. The cross term vanishes because

$$\mathbb{E}\big[(\Lambda - g^\star(X))(g^\star(X) - g(X))\big] = \mathbb{E}\big[\mathbb{E}[\Lambda - g^\star(X) \mid X](g^\star(X) - g(X))\big] = 0. \tag{A.31}$$

The minimizer is therefore unique up to almost-sure equality.

The multiplicative rule in Eq. (12) admits the log-space parameterization

$$\widehat{g}(u, r) = b_0 + \beta \log u + \gamma \log(1 - r). \tag{A.32}$$

For $\gamma > 0$, this expression is defined on the positive support $r < 1$. The projection argument can be applied under the sampling distribution conditioned on $J \in A$, provided the corresponding second moments are finite. Positions with zero weight are handled by the support-set analysis rather than by assigning a finite value to $\log 0$. For $\gamma = 0$, the hidden term is omitted.

Exponentiation gives

$$\exp(\widehat{g}(u_i, r_i)) = e^{b_0} u_i^{\beta} (1 - r_i)^{\gamma}. \tag{A.33}$$

Thus, multiplication of the two factors corresponds to addition in log space and does not require statistical independence of U and R. The implemented weights set $b_0 = 0$. This scale convention matters for unnormalized sums, although a common factor cancels from normalized weights.

The projection defines an ideal target under limited observability. The fixed ReAgg parameterization is not fitted to this target, and the L^2 identity alone does not imply the uniform log-error condition in Eq. (A.27).

B Geometric Interpretation of Hidden Discrepancy

Let $h_{i,L}^{\text{base}}$ and $h_{i,L}^{\text{inst}}$ be nonzero, aligned hidden states in the same feature space. Define their unit-normalized versions as

$$\widehat{h}_i^{\text{base}} = \frac{h_{i,L}^{\text{base}}}{\|h_{i,L}^{\text{base}}\|_2}, \qquad \widehat{h}_i^{\text{inst}} = \frac{h_{i,L}^{\text{inst}}}{\|h_{i,L}^{\text{inst}}\|_2}. \tag{B.1}$$

The discrepancy in Eq. (9) satisfies

$$\begin{aligned} \|\widehat{h}_i^{\text{base}} - \widehat{h}_i^{\text{inst}}\|_2^2 &= 2 - 2\langle \widehat{h}_i^{\text{base}}, \widehat{h}_i^{\text{inst}} \rangle \\ &= 2d_i. \end{aligned} \tag{B.2}$$

Thus, d_i is half the squared distance between normalized representations and lies in $[0, 2]$.

For a common scalar readout g that is K_g-Lipschitz on a domain containing both representations,

$$|g(u) - g(v)| \le K_g \|u - v\|_2, \tag{B.3}$$

we obtain

$$\left| g(\widehat{h}_i^{\text{base}}) - g(\widehat{h}_i^{\text{inst}}) \right| \le K_g \sqrt{2d_i}. \tag{B.4}$$

This is a conditional bound for the same readout applied to two representations. The paired language models need not have identical readout functions, so it is not automatically a bound on their actual token scores. It also does not identify d_i with the conditional variance v_i. The geometric relation motivates using discrepancy as an observable cue; the low-discrepancy weighting direction remains a method design choice.

C Algebraic Properties of the Aggregation Operators

Throughout this section, the valid position set and local evidence are fixed. The weights satisfy $w_i \ge 0$. We suppress the carrier superscript and write $W = \sum_i w_i > 0$ and $a_i = w_i / W$, with sums over $\mathcal{I}_D$. The identities concern the ideal operators; their finite-precision implementation is specified in Appendix D.4.

C.1 Normalized Weighted Mean

The weighted mean in Eq. (13) satisfies

$$S_{\text{mean}}(\mathbf{w}) = \frac{\sum_i w_i z_i}{W} = \sum_i a_i z_i. \tag{C.1}$$

For any scalar $c > 0$, $S_{\text{mean}}(c\mathbf{w}) = S_{\text{mean}}(\mathbf{w})$. Constant weights recover the arithmetic mean. This is the operator directly covered by the conditional variance analysis in Appendix A.

C.2 Shared-Weight Ratio

For a carrier with constant $c_D > 0$, write the ratio in Eq. (14) as

$$R(\mathbf{a}) = \frac{\sum_i a_i n_i}{c_D \sum_i a_i m_i}. \tag{C.2}$$

The numerator and denominator use the same valid positions and weights. The ratio is unchanged by a common positive scaling of the weights.

Let $\mathbf{a}$ and $\mathbf{b}$ be two normalized weight vectors whose weighted denominators are nonzero. Define the residual relative to $\mathbf{b}$ as

$$e_i(\mathbf{b}) = n_i - c_D R(\mathbf{b}) m_i. \tag{C.3}$$

By the definition of $R(\mathbf{b})$, $\sum_i b_i e_i(\mathbf{b}) = 0$. Hence,

$$\begin{aligned} R(\mathbf{a}) - R(\mathbf{b}) &= \frac{\sum_i a_i [n_i - c_D R(\mathbf{b}) m_i]}{c_D \sum_i a_i m_i} \\ &= \frac{\sum_i (a_i - b_i) e_i(\mathbf{b})}{c_D \sum_i a_i m_i}. \end{aligned} \tag{C.4}$$

This exact identity relates score changes to weight changes and baseline residuals. It does not imply that reweighting reduces the variance of a ratio. The weighted denominator also determines the sensitivity of the score to those changes.

C.3 Unnormalized Weighted Sum

The sum operator in Eq. (15) decomposes as

$$S_{\text{sum}}(\mathbf{w}) = \sum_i w_i z_i = W \sum_i a_i z_i = W S_{\text{mean}}(\mathbf{w}). \tag{C.5}$$

Thus, $S_{\text{sum}}(c\mathbf{w}) = cS_{\text{sum}}(\mathbf{w})$ for $c > 0$.

When both W and $\mathbf{a}$ are fixed given C, the model in Eq. (A.1) gives

$$\mathbb{E}[S_{\text{sum}} \mid Y = y, C] = W\mu_y(C), \tag{C.6}$$

$$\text{Var}(S_{\text{sum}} \mid Y = y, C) = W^2 \mathbf{a}^\top \Sigma_y(C)\mathbf{a}. \tag{C.7}$$

Pooling within-class variances yields

$$\sum_y q_y \,\text{Var}(S_{\text{sum}} \mid Y = y, C) = W^2 V(\mathbf{a} \mid C). \tag{C.8}$$

The class-conditional expression uses $\Sigma_y(C)$, whereas the pooled expression uses $\bar{\Sigma}(C)$.

The score therefore depends on both relative allocation and total weight. A reduction in normalized variance alone does not establish an improvement for an unnormalized sum when total weight also changes.

C.4 Weighted Spectral Statistic

Index the valid surprisal sequence in its original order as $s_1, \ldots, s_n$. For positive standard deviation, define

$$\bar{s} = \frac{1}{n}\sum_i s_i, \qquad \sigma_s^2 = \frac{1}{n}\sum_i (s_i - \bar{s})^2, \qquad q_i = \frac{s_i - \bar{s}}{\sigma_s}. \tag{C.9}$$

The weighted mean, normalized window, and windowed sequence are

$$\mu_w = \frac{\sum_i w_i q_i}{W}, \qquad \bar{w} = \frac{W}{n}, \qquad \widetilde{q}_i = \frac{w_i}{\bar{w}}(q_i - \mu_w). \tag{C.10}$$

Their sum is

$$\sum_i \widetilde{q}_i = \frac{1}{\bar{w}}\left(\sum_i w_i q_i - \mu_w W\right) = 0. \tag{C.11}$$

With the unnormalized forward transform

$$F_k = \sum_{i=1}^{n} \widetilde{q}_i \exp\left(-\frac{2\pi \mathrm{i} k(i-1)}{n}\right), \tag{C.12}$$

Eq. (C.11) gives $F_0 = 0$. The score in Eq. (16) is

$$S_{\text{Fourier}} = \frac{1}{n}\sum_{k=1}^{K} |F_k|^2, \qquad K = \min(10, \lfloor n/2 \rfloor). \tag{C.13}$$

The real-input FFT retains the nonnegative-frequency coefficients; no additional doubling of their squared magnitudes is used.

For constant positive weights, $\mu_w = \bar{q} = 0$ and $w_i/\bar{w} = 1$, so $\widetilde{q}_i = q_i$. The operator therefore recovers the unweighted spectral score. In exact arithmetic, a common positive weight scaling also leaves μ_w and $w_i/\bar{w}$ unchanged. Numerical conventions are specified in Appendix D.4.

These identities establish DC cancellation and constant-weight recovery. They do not establish reduced spectral leakage or transfer the linear inverse-variance optimality result to the spectral score.

D CARRIER DEFINITIONS AND IMPLEMENTATION DETAILS

D.1 TOKEN AND HIDDEN-STATE ALIGNMENT

The reference models are `tiiuae/falcon-7b` and `tiiuae/falcon-7b-instruct`. They use a shared tokenizer and compatible vocabularies. Both receive the same token sequence. The extraction manifests specify a maximum input length of 1,024 tokens, a batch size of one, and bfloat16 model computation. Autoregressive shifting leaves at most 1,023 target positions before carrier-specific validity filtering. Scores and weights are recomputed from the frozen feature caches; the parameter grid and permutation controls do not require additional model forward passes. For target token t_i, probabilities and hidden states are taken from the preceding autoregressive position.

In tensor notation, the aligned arrays are

```
target_tokens = input_ids[:, 1:]
target_mask   = attention_mask[:, 1:]
token_logits  = logits[:, :-1, :]
token_hidden  = hidden[:, :-1, :]
```

The hidden-state feature uses the last layer, corresponding to the `last_1` feature group. No layer averaging is used for this feature.

The validity masks exclude padded or otherwise unavailable target positions. The extraction and scoring pipeline checks document identifiers, token identifiers, masks, and sequence lengths for alignment. For DNA-style ratio, the effective mask is the intersection of the trajectory and DNA-component masks.

Positions and discrepancy ranks are computed again on the carrier's effective set $\mathcal{I}_D$. This keeps local evidence, hidden discrepancy, and weights aligned. Valid positions retain their original order, including for spectral scoring.

D.2 LOCAL EVIDENCE DEFINITIONS

Let $\mathcal{V}$ be the shared vocabulary and define

$$P_i^{\text{base}}(v) = p_{\text{base}}(v \mid x_{<i}), \qquad P_i^{\text{inst}}(v) = p_{\text{inst}}(v \mid x_{<i}). \tag{D.1}$$

All logarithms are natural. The following definitions specify carrier evidence and mathematical scores. Raw uses unit weights, while ReAgg changes only the weights passed to the corresponding operator. Score orientation and evaluation protocols are specified separately in Appendix E.2.

Likelihood. The local evidence is the base-model log-probability of the observed target:

$$\ell_i^{\text{base}} = \log P_i^{\text{base}}(t_i). \tag{D.2}$$

Likelihood uses ℓ_i^{base} with the weighted mean operator.

Entropy. The local evidence is the entropy of the base-model next-token distribution:

$$H_i^{\text{base}} = -\sum_{v \in \mathcal{V}} P_i^{\text{base}}(v) \log P_i^{\text{base}}(v). \tag{D.3}$$

Entropy uses H_i^{base} with the weighted mean operator.

Log-Rank. Define the observed token's rank by

$$\rho_i = 1 + \sum_{v \in \mathcal{V}} \mathbf{1}\left\{P_i^{\text{base}}(v) > P_i^{\text{base}}(t_i)\right\}. \tag{D.4}$$

Here, $\mathbf{1}\{\cdot\}$ is an indicator. The implementation counts vocabulary logits strictly greater than the target logit, which gives the same rank. The local evidence is

$$z_i^{\text{LogRank}} = -\log \rho_i. \tag{D.5}$$

Log-Rank applies the weighted mean to this negative log-rank statistic.

LRR. The shared-weight ratio uses

$$n_i^{\mathrm{LRR}} = \ell_i^{\mathrm{base}}, \qquad m_i^{\mathrm{LRR}} = -\log \rho_i, \qquad c_D = 1. \tag{D.6}$$

Thus, its numerator and denominator are the weighted log-likelihood and weighted negative log-rank, respectively.

IRM-mean and IRM-sum. Define

$$\Delta_i = \log P_i^{\mathrm{inst}}(t_i) - \log P_i^{\mathrm{base}}(t_i). \tag{D.7}$$

Both carriers use the same instruction–base log-probability difference Δ_i (Liu et al., 2025). IRM-mean applies normalized weighted averaging, whereas IRM-sum applies the unnormalized weighted sum.

DNA-style ratio. Let

$$\widehat{t}_i = \arg\max_{v \in \mathcal{V}} P_i^{\mathrm{inst}}(v). \tag{D.8}$$

The token-aligned numerator and denominator components are

$$n_i^{\mathrm{DNA}} = -\log P_i^{\mathrm{inst}}(t_i) - \log P_i^{\mathrm{inst}}(\widehat{t}_i), \tag{D.9}$$

$$m_i^{\mathrm{DNA}} = -\sum_{v \in \mathcal{V}} P_i^{\mathrm{base}}(v) \log P_i^{\mathrm{inst}}(v). \tag{D.10}$$

The score is

$$S_{\mathrm{DNA}}(\mathbf{w}) = \frac{\sum_{i \in \mathcal{I}_D} w_i n_i^{\mathrm{DNA}}}{2 \sum_{i \in \mathcal{I}_D} w_i m_i^{\mathrm{DNA}}}. \tag{D.11}$$

The factor of two in the score averages the instruction model's negative log-probabilities of the observed token and its own top-ranked token. The denominator is the cross-entropy from the base distribution to the instruction distribution. Both use the same observed prefix. This defines the token-aligned DNA-style carrier used here, rather than the complete original document-level DNA procedure.

Fourier-style. The input sequence is base-model surprisal,

$$s_i = -\ell_i^{\mathrm{base}}. \tag{D.12}$$

It is standardized before weighting and then processed by the spectral operator in Appendix C.4. The carrier measures energy in the specified low-frequency coefficients; it does not reproduce every component of an external Fourier-based detector.

D.3 Rank Normalization and Weight Construction

Rank normalization in Eq. (10) is performed separately within each carrier's effective set. Let $\varepsilon_{\mathrm{num}} = 10^{-12}$. For finite discrepancy values on a nonempty valid set, the implementation uses

$$r_i = \begin{cases} \frac{1}{2}, & T_D = 1 \text{ or } \max_j d_j - \min_j d_j \le \varepsilon_{\mathrm{num}}, \\ \dfrac{\mathrm{rank}_{\mathcal{I}_D}(d_i) - 1}{T_D - 1}, & \text{otherwise.} \end{cases} \tag{D.13}$$

Ranks are ascending, and exactly tied values receive average ranks. Non-finite discrepancy values are rejected rather than assigned an artificial rank.

The default weights in Eq. (12) are

$$w_i = \left(\frac{\kappa(i)}{T_D}\right)^{\beta} (1 - r_i)^{\gamma}, \qquad \beta = \gamma = 1. \tag{D.14}$$

Setting $\beta = 0$ removes the position factor. At $\gamma = 0$, both hidden-discrepancy branches are defined to equal one, including at rank endpoints. The high-discrepancy control replaces $(1 - r_i)^{\gamma}$ with r_i^{γ}. These conventions recover unit weights when $\beta = \gamma = 0$.

For finite exponents and a nonempty sequence, the default construction has positive total weight in exact arithmetic. If all discrepancies are tied, every hidden factor equals $2^{-\gamma} > 0$. Otherwise, at least one position has $r_i < 1$, and its position factor is strictly positive. Zero weights can occur at $r_i = 1$ when $\gamma > 0$. The implementation does not add a positive floor to these weights. These positions are covered by the support-set analysis in Appendix A.4.

D.4 Numerical Conventions

The scoring implementation uses $\varepsilon_{\mathrm{num}} = 10^{-12}$ for the following checks. Empty effective sequences have undefined scores.

Weighted means and ratios. A weighted mean is invalid when its total weight is at most $\varepsilon_{\mathrm{num}}$. For LRR, the denominator check is applied to the weighted mean:

$$\left| \frac{\sum_i w_i m_i^{\mathrm{LRR}}}{W} \right| \leq \varepsilon_{\mathrm{num}}. \tag{D.15}$$

For DNA-style ratio, the check is instead applied to the weighted denominator sum:

$$\sum_i w_i m_i^{\mathrm{DNA}} \leq \varepsilon_{\mathrm{num}}. \tag{D.16}$$

These cases return an invalid score rather than a clipped finite ratio.

Spectral scoring. Sequences with fewer than four valid positions return an invalid spectral score. Standardization uses the population standard deviation, with divisor T_D. If $\sigma_s \leq \varepsilon_{\mathrm{num}}$, the standardized sequence is set to zero.

For nonconstant weights, the implemented window is

$$\tau_i^{\mathrm{num}} = \frac{w_i}{\bar{w} + \varepsilon_{\mathrm{num}}}, \qquad \widetilde{q}_i^{\mathrm{num}} = \tau_i^{\mathrm{num}}(q_i - \mu_w). \tag{D.17}$$

Relative to the exact operator in Eq. (16),

$$\widetilde{q}_i^{\mathrm{num}} = \alpha \widetilde{q}_i, \qquad \alpha = \frac{\bar{w}}{\bar{w} + \varepsilon_{\mathrm{num}}}. \tag{D.18}$$

Consequently, in exact arithmetic the nonconstant-weight branch scales the ideal spectral score by α^2. The common denominator still preserves the zero-sum identity. Exact scale invariance refers to the mathematical operator; the numerical stabilizer introduces this explicit scale dependence.

The implementation bypasses windowing when the weights are numerically constant, as determined by `numpy.allclose` with its default tolerances. It then applies the FFT directly to the standardized sequence. The forward `numpy.fft.rfft` is unnormalized, the DC coefficient is omitted, and squared magnitudes are summed without one-sided power doubling.

Numerically invalid scores are represented as non-finite values and are not replaced by arbitrary constants. Floating-point computation may leave a small residual DC component even when the corresponding algebraic sum is zero.

E Experimental Protocols and Additional Results

E.1 Benchmark Composition and Evaluation Scope

The evaluation collection contains 18 subsets from DetectRL, M4, EvoBench, MAGE, and RealDet, totaling 24,210 text records. This is a count of records across subsets, not a claim that every text or source is unique across the collection. The 13 primary subsets contain 22,710 records; the five supplementary subsets contain 1,500 records. Table E.1 lists their sizes and splitting units. The identifiers in this table are used in the detailed result tables. All five supplementary subsets use the GPT-4o Latest generation setting identified in their dataset keys.

Primary splits. The primary protocol uses five data seeds, $s \in \{0, 1, 2, 3, 4\}$. For each seed, the implementation randomly permutes the available splitting units, reserves the first $\lfloor G/2 \rfloor$ units for development, and uses the remaining units for evaluation, where G is the number of units. Thus, a 50/50 division of groups need not produce equal document counts. The same splitting procedure is used across carriers and weighting settings. ReAgg Fixed does not fit parameters on the development partition. There is no detector-training partition in this fixed comparison. The term `strict` in the result files denotes this split-based scope; it is not a blanket guarantee of complete source independence.

Table E.1: Benchmark composition. The splitting unit describes the available grouping information; it does not imply that all underlying source relationships are known.

ID	Subset	Records	Scope	Splitting unit
D1	DetectRL MultiDomain	2,000	Primary	Document group
D2	DetectRL MultiLLM	2,000	Primary	Document group
M1	DNA-M4	2,000	Primary	Document group
M2	M4 WikiHow–ChatGPT	2,000	Primary	Pair
M3	M4 WikiHow–Davinci	2,000	Primary	Pair
M4	M4 Wikipedia–ChatGPT	2,000	Primary	Pair
E1	EvoBench Gemini Exp.	1,486	Primary	Domain group
E2	EvoBench Gemini Latest	1,500	Primary	Domain group
E3	EvoBench GPT-4	1,500	Primary	Domain group
E4	EvoBench GPT-4o Latest	1,500	Primary	Domain group
E5	EvoBench Qwen	1,200	Primary	Domain group
G1	MAGE GPT OOD	1,524	Primary	Document
R1	RealDet	2,000	Primary	Document group
S1	EvoBench Harmful	300	Suppl.	Full subset
S2	EvoBench PeerRead	300	Suppl.	Full subset
S3	EvoBench PubMed	300	Suppl.	Full subset
S4	EvoBench Writing	300	Suppl.	Full subset
S5	EvoBench XSum	300	Suppl.	Full subset
Total	18 subsets	24,210	13 + 5	—

Available grouping information. For DetectRL, document groups use recovered source components. DNA-M4 uses exact-duplicate relations and partially known source identities; 1,261 of its 2,000 records have unresolved source identity. The three WikiHow/Wikipedia M4 subsets use verified pair identifiers. For the five primary EvoBench subsets, the grouping field corresponds to *domain*, not individual source documents. These subsets have four or five domain groups, so different seeds can reuse the same domain partitions. MAGE uses document-level random splitting. RealDet groups normalized exact-text duplicates; other source relationships remain unknown.

Supplementary scope and metadata limits. The five single-domain EvoBench subsets are evaluated on their full records without development/test partitioning. Each contributes one full-subset score per deterministic setting, rather than five independent evaluations. These results are reported separately from the primary comparison. The available metadata do not establish complete independence across subsets, generators, or source documents. In particular, source grouping based on known relations cannot exclude unknown shared provenance. The record total must therefore not be interpreted as the size of an independently sampled population of unique documents.

E.2 Metric Definition and Aggregation Order

Let $U_{bDs}^{(m)}$ be the ordinary AUROC for subset b, carrier D, data split s, and weighting setting m, with machine-generated text as the positive class. The reported score is

$$A_{bDs}^{(m)} = \max\big(U_{bDs}^{(m)}, 1 - U_{bDs}^{(m)}\big). \tag{E.1}$$

Each setting chooses its orientation using the labels of its own evaluation split. This is direction-corrected AUROC, denoted AUROC* in tables. It describes ranking separability after orientation selection; it is not performance under a direction fixed before testing. The convention is applied equally to Raw, ReAgg, ablations, and controls. Symmetric treatment does not remove evaluation-label dependence.

For a primary subset, scores are first averaged over the five data seeds:

$$\overline{A}_{bD}^{(m)} = \frac{1}{5}\sum_{s=0}^{4} A_{bDs}^{(m)}. \tag{E.2}$$

For a full-evaluation subset, $\overline{A}_{bD}^{(m)}$ is its single full-subset score. For a collection $\mathcal{B}$ of subsets and $\mathcal{D}$ of eight carriers, the macro-average is

$$\mathrm{Macro}^{(m)}(\mathcal{B}) = \frac{1}{8|\mathcal{B}|} \sum_{b \in \mathcal{B}} \sum_{D \in \mathcal{D}} \overline{A}_{bD}^{(m)}. \tag{E.3}$$

Benchmark-family scores average their constituent subsets equally. The Overall column averages all 13 primary subsets, so it is not an unweighted average of the five benchmark-family columns.

The change in percentage points and positive-combination count are

$$\Delta_{bD} = 100\big(\overline{A}_{bD}^{(\mathrm{ReAgg})} - \overline{A}_{bD}^{(\mathrm{Raw})}\big), \tag{E.4}$$

$$N_+(\mathcal{B}) = \sum_{b \in \mathcal{B}} \sum_{D \in \mathcal{D}} \mathbf{1}\{\Delta_{bD} > 0\}. \tag{E.5}$$

Gains and signs are computed before rounding. Consequently, subtracting two printed scores can differ by 0.01 points from the printed gain. Positive-combination counts describe coverage; combinations sharing texts, sources, or splits are not independent trials.

Finite-score handling. Local validity masks are constructed as described in Appendix D.1. At evaluation time, non-finite document scores are excluded rather than replaced by constants. Each evaluated score vector must retain at least ten finite records and both classes. The primary summaries apply this check to each setting separately. The control comparisons additionally use the intersection of records with finite Raw, ReAgg, and control scores for each comparison. Thus, identical input partitions do not imply identical finite-score counts under every operator and weighting setting. In the archived Raw and ReAgg Fixed summaries, the reported finite counts match for every subset–carrier–split comparison. Finite counts are retained in the result records.

E.3 Weighting Configurations and Permutation Controls

Let $p_i = u_i$ and $\eta_i = 1 - r_i$ on the carrier's effective positions. The fixed configurations use the weights in Table E.2. All active exponents equal one. A permutation π acts within a document's effective position set. For the hidden-factor control, the code permutes discrepancy values and recomputes their ranks; because ranking is permutation-equivariant, this is equivalent to permuting the hidden factors, including tied ranks.

Table E.2: Weighting configurations. Permutations are applied within a document and never exchange tokens between documents.

Setting	Unnormalized token weight
Raw	1
Position-only	p_i
Hidden-low-only	η_i
ReAgg Fixed	$p_i\eta_i$
Joint-high	$p_i r_i$
Joint-weight shuffle	$(p\eta)_{\pi(i)}$
Position shuffle	$p_{\pi(i)}\eta_i$
Hidden shuffle	$p_i\eta_{\pi(i)}$

The permutation seeds are $42, 123, 456, 789, 1024$. For each subset–carrier combination, AUROC* is calculated separately for each control seed and data split, averaged over control seeds and then data splits, and finally macro-averaged across combinations. Predictions are not averaged across control seeds before computing AUROC. The high-discrepancy and uniform settings are deterministic; repeated control-seed entries for these settings do not represent independent runs. The code name `random` refers to joint-weight shuffling, not newly drawn independent random weights.

Joint-weight shuffling preserves the weight multiset, total weight, normalized-weight concentration, and number of zero weights, but changes the positions receiving those weights. Factor-wise shuffling preserves only the shuffled factor's multiset; it can change the products and their total weight. This

distinction matters particularly for IRM-sum, whose score is not invariant to document-dependent changes in total weight.

E.4 Results by Evaluation Scope

Table E.3 separates the primary and supplementary scopes. The all-subset result is an additional descriptive average, not a pooled training result or an independent validation set.

Table E.3: Scope-level direction-corrected AUROC (%). The final column counts subset–carrier combinations with positive unrounded gains.

Scope	Raw	ReAgg	Δ (pp)	Positive
Primary (13 subsets)	75.22	77.54	+2.32	79/104
Supplementary (5 subsets)	70.20	73.55	+3.35	29/40
All subsets (descriptive)	73.82	76.43	+2.61	108/144

E.5 Complete Subset-Level Results

Tables E.4–E.11 report all 18 subsets for each carrier, including both single-cue ablations. The first 13 rows use the primary protocol; the final five use full-subset supplementary evaluation. P-only and H-only denote Position-only and Hidden-low-only, respectively. The final column is ReAgg minus Raw in percentage points. The tables preserve losses as well as improvements.

Table E.4: Likelihood: complete direction-corrected AUROC (%). Subset identifiers are defined in Table E.1.

Subset	Raw	P-only	H-only	ReAgg	Δ (pp)
D1	58.04	53.02	55.71	52.97	−5.07
D2	61.73	65.47	62.49	66.52	+4.79
M1	86.40	87.32	88.47	88.37	+1.97
M2	97.06	96.06	97.41	96.33	−0.73
M3	81.48	79.70	84.43	82.37	+0.88
M4	92.77	92.11	94.49	94.29	+1.51
E1	70.22	73.02	71.88	74.03	+3.81
E2	59.00	61.44	60.38	62.32	+3.32
E3	62.11	63.76	65.90	66.90	+4.79
E4	65.92	67.46	69.88	70.28	+4.36
E5	86.32	89.00	86.69	89.40	+3.07
G1	78.10	81.85	78.32	82.39	+4.29
R1	85.01	86.06	87.46	87.95	+2.94
S1	54.35	55.18	50.52	51.29	−3.06
S2	69.28	77.14	70.75	76.94	+7.66
S3	83.72	80.64	88.81	84.85	+1.13
S4	87.36	86.23	88.43	88.18	+0.81
S5	58.62	65.89	66.56	67.70	+9.08
Primary mean	75.71	76.64	77.19	78.01	+2.30
Suppl. mean	70.67	73.01	73.01	73.79	+3.13

Table E.5: Entropy: complete direction-corrected AUROC (%). Subset identifiers are defined in Table E.1.

Subset	Raw	P-only	H-only	ReAgg	Δ (pp)
D1	74.11	68.89	69.34	64.90	-9.21
D2	53.99	51.60	51.08	55.87	$+1.88$
M1	72.53	77.08	78.71	80.94	$+8.41$
M2	97.06	96.45	97.15	96.59	-0.47
M3	76.74	75.83	80.48	79.02	$+2.29$
M4	76.55	81.23	83.20	87.65	$+11.10$
E1	53.31	53.00	56.96	56.93	$+3.62$
E2	54.70	57.01	52.64	52.56	-2.14
E3	55.16	55.50	56.18	55.17	$+0.01$
E4	55.60	54.21	60.25	58.54	$+2.94$
E5	60.50	65.64	67.14	71.51	$+11.00$
G1	61.79	66.33	66.61	70.97	$+9.18$
R1	75.44	78.74	81.00	83.18	$+7.74$
S1	54.44	56.11	51.34	53.43	-1.01
S2	59.65	58.76	50.95	50.69	-8.96
S3	70.83	66.68	80.62	75.05	$+4.22$
S4	72.64	68.45	78.98	77.32	$+4.68$
S5	69.97	67.17	58.96	60.92	-9.05
Primary mean	66.73	67.81	69.29	70.29	$+3.57$
Suppl. mean	65.51	63.43	64.17	63.48	-2.02

Table E.6: Log-Rank: complete direction-corrected AUROC (%). Subset identifiers are defined in Table E.1.

Subset	Raw	P-only	H-only	ReAgg	Δ (pp)
D1	56.19	52.87	53.68	53.06	−3.13
D2	62.66	67.06	63.64	68.37	+5.71
M1	87.80	88.75	89.99	89.97	+2.17
M2	96.83	95.71	97.14	95.85	−0.98
M3	81.03	79.24	84.24	82.22	+1.19
M4	94.64	94.00	96.36	96.10	+1.46
E1	69.98	73.39	71.42	74.47	+4.48
E2	59.38	62.52	60.43	63.32	+3.94
E3	61.64	63.57	65.41	67.04	+5.40
E4	64.65	66.92	68.95	70.14	+5.50
E5	86.88	89.50	86.94	89.85	+2.98
G1	78.36	82.53	78.34	82.90	+4.54
R1	85.98	86.83	88.20	88.76	+2.78
S1	55.17	55.49	51.36	52.00	−3.17
S2	68.56	76.44	70.28	77.09	+8.53
S3	81.83	78.75	87.12	83.48	+1.65
S4	85.09	84.04	86.70	86.78	+1.69
S5	57.82	68.02	67.84	70.52	+12.70
Primary mean	75.85	77.14	77.29	78.62	+2.77
Suppl. mean	69.69	72.55	72.66	73.98	+4.28

Table E.7: IRM-mean: complete direction-corrected AUROC (%). Subset identifiers are defined in Table E.1.

Subset	Raw	P-only	H-only	ReAgg	Δ (pp)
D1	79.75	76.99	75.99	71.51	−8.23
D2	79.43	78.86	82.99	80.30	+0.87
M1	74.16	77.21	77.93	78.80	+4.64
M2	95.96	91.67	97.80	94.67	−1.29
M3	93.90	90.64	95.34	91.94	−1.96
M4	92.21	91.73	95.47	93.69	+1.48
E1	71.51	75.22	75.32	77.89	+6.37
E2	68.72	72.01	69.30	72.37	+3.65
E3	82.13	85.56	83.23	85.04	+2.92
E4	71.55	76.57	73.67	76.96	+5.41
E5	78.57	82.85	80.69	83.37	+4.80
G1	92.63	91.42	92.16	90.63	−2.00
R1	85.46	82.98	85.48	83.31	−2.15
S1	62.32	63.27	58.33	61.30	−1.02
S2	71.89	78.25	74.26	79.00	+7.12
S3	87.13	91.21	90.75	93.02	+5.89
S4	75.44	77.68	78.98	79.96	+4.52
S5	66.16	78.55	70.82	78.20	+12.04
Primary mean	82.00	82.59	83.49	83.11	+1.12
Suppl. mean	72.59	77.79	74.63	78.30	+5.71

Table E.8: LRR: complete direction-corrected AUROC (%). Subset identifiers are defined in Table E.1.

Subset	Raw	P-only	H-only	ReAgg	Δ (pp)
D1	52.93	59.71	54.50	62.01	+9.08
D2	64.37	70.29	65.92	72.84	+8.47
M1	88.20	89.15	90.00	90.24	+2.04
M2	94.14	91.51	93.16	89.68	−4.46
M3	78.27	76.09	82.06	79.73	+1.45
M4	97.01	96.46	97.39	97.02	+0.01
E1	66.34	71.09	66.39	71.74	+5.40
E2	58.75	63.24	58.91	63.77	+5.01
E3	57.79	60.95	60.23	64.82	+7.04
E4	57.98	62.27	61.60	65.99	+8.01
E5	84.58	87.52	83.29	87.90	+3.32
G1	76.82	81.64	76.04	81.33	+4.51
R1	87.14	86.99	88.55	89.05	+1.92
S1	56.16	56.60	53.65	52.88	−3.29
S2	58.42	65.42	61.08	67.44	+9.01
S3	67.75	69.39	65.64	73.06	+5.31
S4	71.10	73.11	75.11	77.96	+6.86
S5	52.10	68.26	66.23	72.38	+20.28
Primary mean	74.18	76.69	75.23	78.16	+3.99
Suppl. mean	61.11	66.56	64.34	68.74	+7.63

Table E.9: DNA-style ratio: complete direction-corrected AUROC (%). Subset identifiers are defined in Table E.1.

Subset	Raw	P-only	H-only	ReAgg	Δ (pp)
D1	87.28	88.42	86.98	87.66	+0.38
D2	88.77	88.54	87.58	88.24	−0.53
M1	91.61	88.14	90.13	87.47	−4.15
M2	78.71	65.78	85.54	68.96	−9.76
M3	90.19	85.47	91.20	87.35	−2.84
M4	99.42	95.13	99.41	96.50	−2.92
E1	85.90	87.77	84.44	86.69	+0.78
E2	78.53	83.14	75.50	80.93	+2.40
E3	82.70	85.48	82.25	85.00	+2.30
E4	82.67	84.99	81.90	84.60	+1.93
E5	93.43	93.56	92.89	93.16	−0.27
G1	95.03	95.33	94.19	95.04	+0.00
R1	94.17	92.65	93.73	92.52	−1.65
S1	52.20	53.61	52.97	55.50	+3.30
S2	81.92	89.89	79.87	88.38	+6.46
S3	93.11	94.90	90.91	93.47	+0.36
S4	96.25	97.32	94.83	96.49	+0.24
S5	96.69	97.26	96.77	97.48	+0.80
Primary mean	88.34	87.26	88.13	87.24	−1.10
Suppl. mean	84.04	86.60	83.07	86.27	+2.23

Table E.10: IRM-sum: complete direction-corrected AUROC (%). Subset identifiers are defined in Table E.1.

Subset	Raw	P-only	H-only	ReAgg	Δ (pp)
D1	91.36	90.05	89.10	86.35	−5.01
D2	83.09	82.13	86.17	83.88	+0.79
M1	71.28	73.11	74.86	74.97	+3.69
M2	77.62	72.59	84.71	78.67	+1.05
M3	81.81	79.07	85.46	81.70	−0.12
M4	95.55	95.25	97.13	96.64	+1.08
E1	67.34	71.40	71.72	74.26	+6.92
E2	64.55	68.78	66.45	69.05	+4.51
E3	79.21	83.09	81.45	82.88	+3.67
E4	65.86	70.64	69.16	71.44	+5.58
E5	75.01	79.81	77.99	79.90	+4.89
G1	80.50	80.83	83.28	82.83	+2.33
R1	77.73	74.79	81.36	77.98	+0.25
S1	63.39	64.31	60.26	62.91	−0.48
S2	76.21	81.18	77.59	80.38	+4.16
S3	85.31	89.95	89.86	91.63	+6.32
S4	70.53	73.57	74.90	75.18	+4.64
S5	74.46	83.67	77.16	80.74	+6.28
Primary mean	77.76	78.58	80.68	80.04	+2.28
Suppl. mean	73.98	78.54	75.95	78.17	+4.19

Table E.11: Fourier-style: complete direction-corrected AUROC (%). Subset identifiers are defined in Table E.1.

Subset	Raw	P-only	H-only	ReAgg	Δ (pp)
D1	56.24	65.12	52.74	65.68	+9.44
D2	59.50	61.91	55.93	65.37	+5.87
M1	66.61	51.73	51.11	61.91	−4.70
M2	58.74	64.05	57.97	54.25	−4.49
M3	51.66	54.22	69.98	60.15	+8.49
M4	63.24	79.83	70.13	81.31	+18.07
E1	53.98	65.54	54.49	62.70	+8.72
E2	55.98	58.82	55.81	57.02	+1.04
E3	55.02	63.20	53.02	65.86	+10.84
E4	59.02	64.37	55.90	65.59	+6.58
E5	82.54	52.63	71.97	62.04	−20.49
G1	61.97	63.72	51.44	68.81	+6.83
R1	70.64	60.80	55.87	72.01	+1.37
S1	63.64	57.40	60.00	58.04	−5.60
S2	51.56	71.90	51.44	67.57	+16.01
S3	71.30	58.13	70.33	64.49	−6.81
S4	71.56	59.84	67.47	64.37	−7.19
S5	62.10	74.70	62.08	74.11	+12.01
Primary mean	61.16	62.00	58.18	64.82	+3.66
Suppl. mean	64.03	64.40	62.26	65.72	+1.68

E.6 Carrier-Level Control Comparisons

Table E.12 reports ReAgg minus each control, averaged across the 13 primary subsets using the common finite-score comparisons. Positive values favor ReAgg. The final row reports positive subset–carrier combinations; it is not a significance count. Small differences from subtracting separately rounded macro-averages can arise from both rounding and finite-score intersections.

Table E.12: ReAgg gains over controls (percentage points). J-shuffle, P-shuffle, and H-shuffle denote joint-weight, position-factor, and hidden-factor shuffling.

Carrier	J-shuffle	P-shuffle	H-shuffle	Joint-high
Likelihood	+3.24	+1.32	+1.89	+5.09
Entropy	+3.84	+1.17	+2.59	+4.62
Log-Rank	+3.72	+1.78	+2.02	+5.15
IRM-mean	+4.10	+1.32	+2.31	+4.34
LRR	+5.50	+3.68	+2.34	+5.12
DNA-style ratio	+1.13	+0.30	+1.15	+1.64
IRM-sum	+4.19	+0.10	+2.24	+4.07
Fourier-style	+8.33	+8.01	+4.03	+4.91
Macro-average	+4.26	+2.21	+2.32	+4.37
Positive combinations	87/104	71/104	94/104	90/104

E.7 Sensitivity to the Weighting Exponents

We examine the low-discrepancy joint family on the recorded grid $\beta \in \{0.25, 0.5, 1, 2\}$ and $\gamma \in \{0.5, 1, 2, 4\}$. Table E.13 reports the macro-average gain over Raw on the 13 primary subsets, with the same data splits and AUROC* convention used in the main comparison. These entries are evaluations of fixed grid configurations, not scores obtained by choosing a separate best configuration for each subset.

Table E.13: Sensitivity of the joint low-discrepancy weighting on the primary subsets. Entries are macro-average AUROC* gains in percentage points. Bold marks the predefined default, not the best grid value.

$\beta\backslash\gamma$	0.5	1	2	4
0.25	+1.10	+1.51	+1.55	+0.71
0.5	+1.60	+2.02	+2.15	+1.42
1	+1.98	**+2.32**	+2.40	+1.64
2	+1.70	+1.98	+1.97	+1.09

All 16 configurations have positive macro-average gains on this grid, but this does not imply positive gains for every subset–carrier combination. The default $(\beta, \gamma) = (1, 1)$ gains 2.32 points. The largest grid-average gain is 2.40 points at $(1, 2)$, approximately 0.08 points above the default. Larger γ is not uniformly beneficial: at $\beta = 1$, increasing γ from 1 to 4 reduces the gain to 1.64 points. The default is therefore effective within this tested grid, but is not claimed to be the optimal parameter choice. No statistical equivalence or universal parameter robustness follows from this descriptive comparison.

E.8 Supplementary Threshold-Based Metrics

Table E.14 reports the stored maximum F1 and TPR at an empirical FPR of at most 5% for the primary subsets. For each evaluation split, the score direction is selected as in Eq. equation E.1. Maximum F1 selects the best score threshold using that split's labels. The TPR statistic takes the largest empirical ROC true-positive rate among points with FPR at most 0.05, without interpolation. Values are averaged over splits and then subsets. These are evaluation-selected descriptive metrics, not performance under thresholds calibrated on a separate development set.

Table E.14: Supplementary maximum F1 and TPR at empirical FPR $\leq 5\%$ (all values in %). Directions and operating points use evaluation labels, as described in the text.

Carrier	F1 Raw	F1 ReAgg	TPR Raw	TPR ReAgg
Likelihood	75.65	77.20	37.34	42.05
Entropy	71.27	73.19	22.33	28.30
Log-Rank	75.88	77.47	39.56	44.07
IRM-mean	77.23	77.86	50.39	49.74
LRR	75.09	76.28	40.44	43.41
DNA-style ratio	83.40	82.46	66.82	64.92
IRM-sum	75.42	76.02	36.50	40.92
Fourier-style	68.16	68.70	14.41	15.54
Macro-average	75.26	76.15	38.47	41.12

E.9 Resampling Diagnostics and Interpretation

The archived control diagnostics use paired resampling separately within each data split. For a given subset, carrier, and control, the eligible records are the intersection of finite ReAgg scores and finite scores for all five control seeds. Each of 1,000 attempted bootstrap draws selects one control seed uniformly and resamples the splitting units with replacement; both score vectors are evaluated on the same sampled records. Units are verified pairs for pair-based splits, domains for primary EvoBench, recovered document groups for document-group splits, and individual documents for random or full-subset evaluation. Draws containing only one class are skipped, so the accepted draw count can be smaller than 1,000. The metric is reoriented using the bootstrap sample's labels for each method, consistently with the direction-corrected convention. The stored endpoints are the 2.5th and 97.5th percentiles of the accepted ReAgg-minus-control differences.

The archived file `controls_ci_per_split.csv` contains 2,240 such intervals: $13 \times 8 \times 4 \times 5 = 2{,}080$ for the primary scope and $5 \times 8 \times 4 = 160$ for the supplementary scope. They characterize per-split resampling variation, including the selected control-seed variation; they are not intervals for the macro-average across all subsets and splits. In particular, resampling only four or five domain groups gives a limited basis for inference about unseen domains. Intervals are not averaged, treated as independent replications, or used to turn the positive-combination counts into significance counts. No multiple-comparison correction or family-wide significance claim is made from these diagnostics.

E.10 Reproducibility and Scope of the Evidence

The result record retains each subset's splitting semantics, source run, executed-code hashes, and result-file hashes. The current collection combines compatible outputs from the revised source-group evaluations and the remaining frozen evaluations; it is not represented as originating from a single clean code commit. Detailed tables in this appendix are computed from the same frozen subset–carrier summaries used in the main text. The source record contains both fixed-setting results and exploratory selection outputs. Only the fixed settings specified here are used in the comparisons above; development-selected or evaluation-selected best-in-grid scores are not mixed into the ReAgg Fixed results.

The reported experiments use one frozen base/instruction model pair. They do not establish invariance to the choice of reference models. No additional training refers to the weight construction and aggregation layer, not to the absence of inference cost: hidden-discrepancy features require outputs from both reference models. These appendices make no measured latency, throughput, or memory claim. Likewise, the conditional variance analysis does not certify the empirical proxy as an estimator of oracle precision, and direction-corrected AUROC does not replace evaluation with directions fixed before testing.

E.11 Evaluation with Directions Fixed on the Development Split

We retain the 13 primary subsets, eight carriers, and five original splits, with ReAgg weights fixed at $\beta = \gamma = 1$ on the low-discrepancy branch. For each split and carrier, Raw and ReAgg separately select a score direction using their common finite-score development records. The selected directions are then fixed for evaluation on the common test records. Test AUROC is not reoriented using test labels, and values below 0.5 are retained.

Table E.15 reports the resulting ordinary test AUROC. The macro-average increases from 74.15% to 77.02%, a gain of 2.87 percentage points; 78 of the 104 subset–carrier combinations have positive mean gains. These results use a different orientation protocol from the historical direction-corrected comparisons and are reported separately.

Table E.15: Test AUROC (%) with score directions selected on development records and frozen for testing. Results average five splits within each subset, then average across the 13 primary subsets. Gains use unrounded results. Positive counts refer to subset-level mean gains and are not significance counts.

Carrier	Raw	ReAgg	Δ (pp)	Positive
Likelihood	74.43	77.55	+3.13	11/13
Entropy	64.22	67.90	+3.67	10/13
Log-Rank	74.55	78.53	+3.98	11/13
IRM-mean	82.00	83.11	+1.12	8/13
LRR	72.30	77.35	+5.05	12/13
DNA-style ratio	88.34	87.24	−1.10	6/13
IRM-sum	77.76	80.04	+2.28	11/13
Fourier-style	59.59	64.45	+4.85	9/13
Macro-average	74.15	77.02	+2.87	78/104

Paired uncertainty analysis. For each of the 520 test comparisons, we compute a paired cluster percentile bootstrap interval from 2,000 resamples, retaining the same development-selected directions throughout. Raw and ReAgg use identical sampled units and multiplicities in each resample. The units follow the evaluation protocol: verified pairs, domain groups, document-source groups, or individual documents. The 2.5th and 97.5th percentiles describe conditional uncertainty in the ReAgg-minus-Raw AUROC difference for that split. These intervals are supplied in the per-split result records; they are not pooled into a confidence interval for the macro-average. The EvoBench test partitions contain only two or three domain groups, which limits inference about unseen domains.

Fixing directions on development data addresses evaluation-time orientation selection. It does not remove earlier method-development exposure to some benchmark sources, and the intervals do not cover uncertainty in those research design choices.

F TOKEN-LEVEL CASE STUDY DEFINITIONS

F.1 CASE IDENTITY AND EVALUATION SCOPE

Figure 4 in the main text visualizes the three illustrative documents listed in Table F.1. The outcome labels refer to the corresponding carrier–subset result, not a performance estimate for the displayed document. The two EvoBench-XSum cases belong to supplementary subset S5. They illustrate weight allocation alongside contrasting subset-level outcomes; they are not a random sample of all successful or failed documents.

Table F.1: Case identity and subset-level direction-corrected AUROC (%). Document indices identify records in the exported case data; they are not cross-dataset global identifiers.

Case / carrier	Subset	Doc. index	Raw	ReAgg	Δ (pp)
Improvement / LRR	S5	4	52.10	72.38	+20.28
Degradation / Entropy	D1	619	74.11	64.90	−9.21
Limited gain / DNA	S5	32	96.69	97.48	+0.80

F.2 LOCAL-SIGNAL PROXIES

The plotted signal is a visualization proxy s_i^{vis}:

$$s_i^{\mathrm{vis}} = \begin{cases} \ell_i^{\mathrm{base}}/(-\log \rho_i), & \text{LRR}, \\ H_i^{\mathrm{base}}, & \text{Entropy}, \\ n_i^{\mathrm{DNA}}/(2m_i^{\mathrm{DNA}}), & \text{DNA-style ratio}. \end{cases} \tag{F.1}$$

A pointwise ratio is undefined in the exported data when the absolute value of its denominator is below 10^{-12}. In particular, rank-one targets have zero negative log-rank and can produce undefined LRR proxies. This does not imply that the document-level LRR score is undefined: that score divides aggregated components, as in Appendix C.2, rather than averaging pointwise ratios. Neither ratio proxy is an additive contribution to the actual document score.

For the finite proxy values within each document, the exporter computes

$$m_s = \mathrm{median}_i\, s_i^{\mathrm{vis}}, \qquad d_s = 1.4826\, \mathrm{median}_i\, |s_i^{\mathrm{vis}} - m_s|. \tag{F.2}$$

When d_s is finite and at least 10^{-12}, the standardized proxy is $(s_i^{\mathrm{vis}} - m_s)/d_s$. If this scale is degenerate, the exporter uses the finite-value mean and population standard deviation; if that scale is also degenerate, it returns zeros. The exporter clips the standardized proxy to $[-3, 3]$. The final figure reads these stored values without another standardization, additional clipping, interpolation, or filling of undefined proxies. The stored CSV does not retain the pre-clipping magnitudes.

F.3 POSITION AND WEIGHT DISPLAY

For $n = T_D > 1$, the display uses evenly spaced positions

$$\xi_i = \frac{\kappa(i) - 1}{n - 1}, \qquad \omega_i = n\, \frac{w_i}{\sum_j w_j}. \tag{F.3}$$

The display coordinate ξ_i spans $[0, 1]$ and differs from the position factor $u_i = \kappa(i)/n$ used to construct the weights. It is a visualization coordinate only; it does not replace the scoring position convention. The color value ω_i is the relative ReAgg weight, with $\omega_i = 1$ corresponding to uniform normalized weighting. Each document's weights are normalized over the original effective sequence before any undefined visualization proxies are omitted. The remaining plotted points are not renormalized.

Table F.2 records the displayed coverage. The three color scales are shared and cover $[0, 3.6]$; the largest source relative weight is approximately 3.512, so no source color value is clipped. Every plotted point has the same marker size.

Table F.2: Coverage of the token-level visualization. Undefined proxies are not replaced with zero and do not change the normalization used for the color values.

Case	Effective positions	Plotted	Undefined
LRR / EvoBench-XSum	202	100	102
Entropy / DetectRL-MultiDomain	137	137	0
DNA / EvoBench-XSum	190	190	0

F.4 Interpretation Limits

The plots show where weights are allocated relative to a standardized signal proxy. Colors encode assigned weights, not measured information content or true conditional precision. The missing LRR proxies mean that panel (a) does not display the complete local-signal sequence, even though its color normalization uses all 202 positions. The clipped values also do not preserve the magnitudes of extreme proxy values beyond the display range. A single document cannot yield an AUROC estimate or establish why the carrier improves or degrades on a population of documents. The subset-level outcomes and the conditional theory therefore remain distinct from these qualitative visualizations.