ZS-MTR-RR: TRAINING-FREE LENGTH-CONDITIONED RELIABILITY ROUTING OF GLOBAL AND MULTI-TRAJECTORY EVIDENCE FOR ZERO-SHOT AI-GENERATED TEXT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models increasingly produce fluent, human-like text, posing severe challenges to academic integrity, misinformation governance, and content provenance. Zero-shot detectors avoid collecting labeled data for unseen generators, yet existing approaches face a dilemma: document-level statistics compress token probabilities into a single global score, overlooking positional sequence structure; while trajectory-based methods retain positional information but rely on fixed aggregation, failing to adapt the relative contributions of global and trajectory evidence as text length varies. To address these limitations, we propose ZS-MTR-RR, a training-free framework featuring Zero-Shot Multi-Trajectory Representation with Length-Conditioned Reliability Routing. Using paired frozen base and instruction-tuned models, ZS-MTR-RR extracts aligned token log-probability outputs to construct both a full-document proxy-discrepancy path and a multi-trajectory evidence path capturing late-stage stability, proxy tension, and token-reuse structures. Conditioned on scored length as an indicator of available positional support, a deterministic soft gate balances the two paths, converting fixed signal combination into length-conditioned reliability routing without training an additional detector or learned router. We evaluate ZS-MTR-RR on the five-domain EvoBench benchmark and four RealBench-derived transformations. Under the unified EvoBench protocol, ZS-MTR-RR achieves the highest cross-domain average AUROC (reaching 0.8753) among evaluated methods. Ablation and length-stratified analyses show that the two evidence paths provide complementary discriminative signals and support the benefit of length-conditioned routing over fixed fusion on the primary benchmark.

1 INTRODUCTION

Large language models can now produce fluent and stylistically diverse text at low cost, raising new challenges for academic integrity, misinformation, and source attribution Wu et al. (2025); Dugan et al. (2024); Yu et al. (2025). AI-generated text detection aims to distinguish texts generated by language models from those written by humans. As generators and application domains continue to evolve, retraining a detector for every new setting is impractical. Zero-shot detection has therefore become an important direction.

AI-text detection methods broadly follow two paradigms: supervised and zero-shot detection Yu et al. (2025). **Supervised detectors** learn decision boundaries from labeled human-written and machine-generated texts, often by fine-tuning pretrained encoders. By contrast, **zero-shot detectors** derive document-level scores from frozen proxy language models without fitting a target-domain classifier. Within the zero-shot paradigm, one line of work constructs a single document-level score by aggregating token-level statistical signals across scored positions. Likelihood- and rank-based detectors use observed-token probabilities or ranks; DetectGPT and Fast-DetectGPT analyze probability curvature; and Binoculars and Alignment Imprint compare probability or preference responses from related model pairs Gehrmann et al. (2019); Su et al. (2023); Mitchell et al. (2023); Bao et al. (2024); Hans et al. (2024); Wu et al. (2026). A second line preserves token-level probability sig-

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.

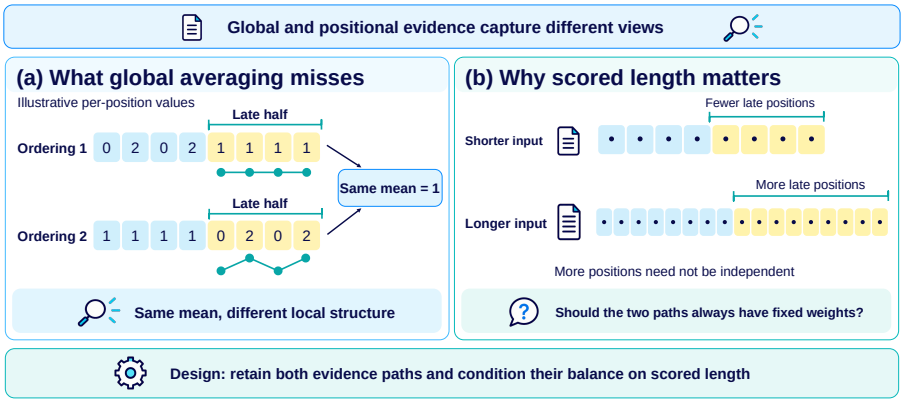

Figure 1: Conceptual motivation for ZS-MTR-RR. (a) Global averaging maps distinct position-wise trajectories to the same score, discarding their order. (b) The effective statistical support of trajectory statistics varies with scored length.

nals as position-indexed trajectories for structural analysis. Lastde examines time-domain structure and SpecDetect analyzes frequency-domain structure, whereas Late-Stage Stability, Telescope, and Temporal Discrepancy Tomography (TDT) focus on late-stage volatility, repetition probability, and time-scale discrepancy patterns, respectively Xu et al. (2025); Luo et al. (2026); Sun et al. (2026b); Nassif & Cooper (2026); West et al. (2025).

These zero-shot methods primarily extract detection evidence through global aggregation or position-sensitive trajectory statistics. As shown in Figure 1(a), global aggregation across all scored positions can map trajectories with different token order and local variation to the same mean, thereby discarding position-sensitive evidence. As shown in Figure 1(b), trajectory statistics preserve positional structure, but the finite statistical support available for estimating local volatility and repetition changes with scored length. Short texts provide fewer scored positions and can yield less stable estimates. Longer texts offer richer trajectory structure, but additional positions need not contribute proportionally more independent information. This length dependence is consistent with empirical and sample-complexity analyses of AI-text detection Ippolito et al. (2020); Chakraborty et al. (2024). This motivates a testable hypothesis: global evidence provides stronger discriminative capacity on short texts, whereas trajectory evidence achieves greater discriminative advantage on longer passages. Based on this hypothesis, we use scored length as a conditioning variable for reliability routing, modulating the contributions of the two evidence paths via a deterministic gate without training an additional router.

To address these issues, we propose **Z**ero-**S**hot **M**ulti-**T**rajectory **R**epresentation with Length-Conditioned **R**eliability **R**outing (**ZS-MTR-RR**), which combines dual-path evidence extraction with length-conditioned reliability routing. The dual-path module captures complementary document-level and position-sensitive evidence from a paired frozen base model and its instruction-tuned counterpart. Instruction tuning alters the conditional token distribution of the proxy, making the likelihood discrepancy between paired models an informative global signal. ZS-MTR-RR combines this global discrepancy with position-sensitive trajectory evidence rather than using it as a standalone detector. ZS-MTR-RR extracts aligned token log-probabilities to form two paths in parallel: a global path aggregating full-document likelihood discrepancy, and a trajectory path combining late-stage stability, proxy tension, and token reuse. The routing module then uses scored length as a routing signal: a deterministic soft gate places higher weight on global evidence for shorter texts and smoothly shifts weight toward trajectory evidence as sequence length increases. This design requires neither an additional detector nor a learned router.

Our main contributions are as follows:

- **A dual-path evidence framework.** We establish a detection framework based on a single frozen base and instruction-tuned proxy pair, unifying full-document proxy divergence with position-sensitive trajectory statistics—including late-stage stability, proxy tension, and token recurrence—

into complementary evidence that preserves both global distributional shifts and local sequential structure without additional model fitting.

- **Training-free length-conditioned reliability routing.** We propose a deterministic soft gate driven by scored length that continuously adjusts the relative contributions of global and trajectory evidence according to available statistical support, achieving adaptive evidence fusion without training an additional detector or learned router.
- **Length-dependent complementarity and routing gains.** Through component ablations, natural-length stratification, source-disjoint diagnostics, proxy sensitivity, and reference-pool transfer, we systematically reveal the complementary behavior of global and trajectory evidence across varying text lengths, validating the performance advantages and operational boundaries of length-conditioned routing over fixed fusion.

2 Related Work

2.1 Supervised AI-Text Detection

Supervised detectors formulate AI-text detection as a binary classification task and learn decision boundaries from labeled human-written and machine-generated texts. Early work fine-tuned RoBERTa to distinguish human text from outputs of GPT-2 variants Solaiman et al. (2019). MAGE evaluates fine-tuned detectors across multiple domains and generators under both in-distribution and transfer settings Li et al. (2024). RADAR improves robustness to paraphrasing by adversarially co-training a detector and a paraphraser Hu et al. (2023). These systems can perform strongly when training and test distributions align, but their reliability depends on the coverage of labeled training data. When evaluation conditions extend beyond that distribution, particularly to unseen generators, domains, or text transformations, generalization performance can drop substantially.

2.2 Zero-Shot Statistical and Model-Pair Detection

Zero-shot detectors derive detection scores from frozen proxy language models without training a task-specific classifier. Likelihood- and rank-based methods build on the tendency of tokens in machine-generated text to receive higher probabilities or better ranks under the chosen proxy model Gehrmann et al. (2019); Su et al. (2023). DetectGPT estimates probability curvature from perturbed alternatives, whereas Fast-DetectGPT estimates conditional probability curvature more efficiently Mitchell et al. (2023); Bao et al. (2024). Model-pair detectors instead exploit statistical discrepancies between related language models. Binoculars combines perplexity and cross-perplexity from base and instruction-tuned models Hans et al. (2024), IRM uses the log-likelihood difference between instruction-tuned and base models as an implicit reward signal Liu et al. (2025), and Alignment Imprint measures preference discrepancies between base and aligned models Wu et al. (2026). Further extensions analyze local-normalization distortion (TempTest) or multiscale uncertainty among low-probability tokens (Uncertainty++) Kempton et al. (2025); Guo et al. (2026). However, these methods ultimately summarize their evidence in a document-level score without explicitly retaining position-indexed trajectory structure.

2.3 Position-Aware Evidence and Adaptive Routing

Position-aware detectors preserve token order before aggregation. Lastde combines temporal statistics with diversity measures, SpecDetect applies spectral analysis, and TDT uses continuous wavelet transforms Xu et al. (2025); Luo et al. (2026); West et al. (2025). Late-Stage Stability tracks derivative volatility Sun et al. (2026b), while Telescope measures token repetition Nassif & Cooper (2026). Similarly, TOCSIN Ma & Wang (2024) incorporates token cohesiveness on long texts, but relies on costly deletion perturbations rather than length-conditioned routing.

A separate research direction explores input-adaptive model selection. DetectRouter learns input–proxy affinities via two-stage prototype-based training, routing each text to a specialized external proxy model Sun et al. (2026a). DetectRouter uses a learned router with source-model supervision and score-based alignment to select among multiple proxy models. In contrast, coordinating document-level and position-aware evidence in a strictly training-free, single-proxy-pair setting remains unresolved. ZS-MTR-RR addresses this gap by using scored length as a continuous routing

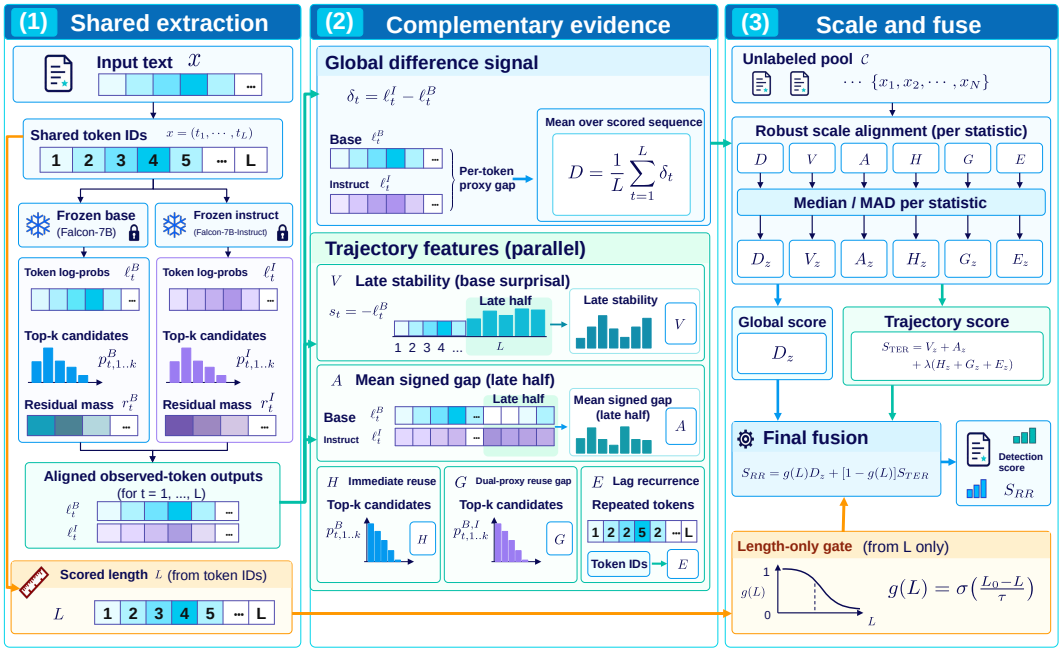

Figure 2: Architecture of ZS-MTR-RR across three stages: (1) Shared extraction from frozen base and instruction-tuned Falcon-7B proxies; (2) Parallel extraction of global discrepancy $D(x)$ and multi-trajectory evidence; (3) Robust scale alignment yielding D_z and S_{TER}, fused via a length-conditioned soft gate into final score S_{RR}. In the diagram, $x = (t_1, \ldots, t_L)$ denotes the scored sequence with prompt context x_0 omitted. Here $\lambda = 0.2$ as in Eq. (8), and $\sigma(\cdot)$ denotes the standard sigmoid gate in Eq. (9).

variable, dynamically balancing global proxy divergence and multi-trajectory evidence via an interpretable closed-form soft gate without training overhead.

3 Method

We present ZS-MTR-RR, a zero-shot detector that combines global dual-proxy discrepancy with multi-trajectory evidence through length-conditioned reliability routing. As illustrated in Figure 2, two frozen proxies first perform shared scoring on the same token sequence; subsequently, we construct in parallel the full-document proxy discrepancy alongside trajectory statistics capturing late-stage stability, proxy tension, and token reuse; finally, an unlabeled reference pool provides robust scale alignment, and a length-conditioned soft gate smoothly balances the contributions of the two evidence paths according to scored length.

3.1 Paired Proxies and Shared Extraction

Given a tokenized text $x = (x_0, x_1, \ldots, x_L)$, the initial token x_0 provides the initial context, while the remaining L tokens undergo probability scoring. We refer to L as the scored length. The base proxy B and instruction-tuned proxy I employ Falcon-7B and Falcon-7B-Instruct, respectively; both models receive identical token IDs and remain frozen throughout evaluation.

For proxy $m \in \{B, I\}$, we define the conditional token log-probability, the base-proxy surprisal, and the cross-proxy discrepancy as:

$$\ell_t^m = \log p_m(x_t \mid x_{<t}), \qquad s_t^B = -\ell_t^B, \qquad \delta_t = \ell_t^I - \ell_t^B. \tag{1}$$

Here, $\delta_t > 0$ indicates that the instruction-tuned proxy assigns a higher probability to the observed token than the base model. As shown in Stage 01 of Figure 2, each frozen proxy performs one forward pass over the same token sequence to obtain the conditional log-probability of each observed

token given its preceding context. The resulting outputs are shared by both evidence paths without generating alternative texts or perturbing the input.

3.2 Global and Multi-Trajectory Evidence

As depicted in Stage 02 of Figure 2, the two evidence paths preserve document-wide average discrepancy and position-sensitive sequence structure, respectively.

Global Proxy Discrepancy. Drawing on cross-model likelihood contrasts Hans et al. (2024); Liu et al. (2025), the global path averages the cross-proxy discrepancy across all scored positions:

$$D(x) = \frac{1}{L}\sum_{t=1}^{L}\delta_t. \tag{2}$$

This score summarizes cross-proxy preference across the full sequence, but discards positional variations. The trajectory path complements this global view across three dimensions: local volatility, late-stage discrepancy, and token reuse.

Late-Stage Stability and Proxy Tension. Following observed late-stage generation stability Sun et al. (2026b), we aggregate local surprisal volatility over the latter half of the scored sequence. Let $\mathcal{I}_{\text{late}} = \{\lfloor L/2 \rfloor + 1, \ldots, L\}$, and define the local window $W_t = \{i : 1 \le i \le L, |i - t| \le 10\}$. Late-stage stability $V(x)$ and late-stage proxy tension $A(x)$ are:

$$V(x) = -\frac{1}{|\mathcal{I}_{\text{late}}|}\sum_{t\in\mathcal{I}_{\text{late}}}\text{sd}\{s_i^B : i \in W_t\}, \qquad A(x) = \frac{1}{|\mathcal{I}_{\text{late}}|}\sum_{t\in\mathcal{I}_{\text{late}}}\delta_t. \tag{3}$$

Here, sd denotes the population standard deviation; larger $V(x)$ indicates a smoother late-stage trajectory. We note that $A(x)$ and the global score $D(x)$ originate from the same proxy discrepancy sequence δ_t, but summarize the latter half and the full sequence, respectively.

Token Reuse Structures. Building upon continuation probability signals Nassif & Cooper (2026), we extract three structural reuse statistics: immediate reuse hits, dual-proxy reuse discrepancy, and multi-lag recurrence.

Let $\mathcal{T}_{t,\kappa}^m$ denote the top-κ candidate set predicted by proxy m under context $x_{\le t}$, with $\kappa = 100$. The negative immediate reuse hit-rate is:

$$H(x) = -\frac{1}{L-1}\sum_{t=1}^{L-1}\mathbf{1}[x_t \in \mathcal{T}_{t,\kappa}^B]. \tag{4}$$

This statistic evaluates whether the model tends to immediately regenerate token x_t after observing it; larger $H(x)$ indicates a lower immediate-reuse hit rate.

To compare reuse tendencies across the proxy pair, let $\widehat{q}_t^m$ denote the estimated probability of regenerating x_t under context $x_{\le t}$. We use model probabilities within top-κ and allocate residual probability mass uniformly outside the set. Restricting evaluation to the latter half $\mathcal{I}_{\text{rep}} = \{\lfloor (L-1)/2 \rfloor + 1, \ldots, L-1\}$, we define:

$$G(x) = \frac{1}{|\mathcal{I}_{\text{rep}}|}\sum_{t\in\mathcal{I}_{\text{rep}}}\left(\log \widehat{q}_t^B - \log \widehat{q}_t^I\right). \tag{5}$$

Thus, larger $G(x)$ indicates that the base proxy assigns higher immediate reuse probability relative to the instruction-tuned model across all late-stage positions rather than conditioning exclusively on observed adjacent duplicates.

Finally, let $K_r = \min(32, L-1)$. We capture lexical repetition patterns across varying lag distances:

$$a_k = \frac{1}{L-k}\sum_{t=1}^{L-k}\mathbf{1}[x_t = x_{t+k}], \qquad \pi_k = \frac{a_k}{\sum_{j=1}^{K_r} a_j}, \qquad E(x) = \sum_{k=1}^{K_r}\pi_k \log(\pi_k + \varepsilon), \tag{6}$$

where $\varepsilon = 10^{-8}$, and $E(x)$ is the negative entropy of the normalized lag distribution; higher values indicate that recurrence matches concentrate at fewer lag distances. When total matching mass satisfies $\sum a_j \le \varepsilon$, this statistic is set to 0.0 before alignment. Tail-mass estimation and short-sequence boundary details are given in Appendix A.

3.3 Robust Scale Alignment and Dual-Path Scores

The six raw statistics differ in location and scale. Following Stage 03 of Figure 2, we standardize each metric via median m_u and MAD d_u over an unlabeled reference pool $\mathcal{C}$:

$$u_z = R_{\mathcal{C}}(u(x)) = \frac{u(x) - m_u}{\widetilde{d}_u}, \qquad u \in \{V, A, H, G, E, D\}, \tag{7}$$

where $\widetilde{d}_u = d_u$ if $d_u \geq \varepsilon$, and $\widetilde{d}_u = 1.0$ otherwise ($\varepsilon = 10^{-8}$). These scale statistics are estimated without ground-truth class labels; reference pool configurations are detailed in Section 4.

The standardized statistics are synthesized into two paths:

$$S_{\mathrm{TER}}(x) = V_z + A_z + 0.2(H_z + G_z + E_z), \qquad D_z(x) = R_{\mathcal{C}}(D(x)). \tag{8}$$

The trajectory path pairs stability and late-stage proxy tension with auxiliary reuse features; the global path preserves full-document discrepancy. Both outputs serve as relative ranking scores with fixed coefficients, requiring no weighting model.

3.4 Length-Conditioned Reliability Routing

The two paths aggregate sequence evidence differently: the global path averages pointwise cross-model differences across all positions, whereas the trajectory path relies on late-stage selection, sliding windows, and recurrence relationships. Consequently, available scored length alters the positional statistical support underlying these metrics. We treat scored length as an observable conditioning variable to dynamically balance the relative contributions of global and trajectory evidence.

As illustrated in Stage 03 of Figure 2, the length gate takes the scored length L obtained from the shared extraction stage and computes the final detection score:

$$g(L) = \frac{1}{1 + \exp\big((L - L_0)/\tau\big)}, \qquad S_{\mathrm{RR}}(x) = g(L)D_z(x) + \big[1 - g(L)\big]S_{\mathrm{TER}}(x), \tag{9}$$

where $g(L) \in (0, 1)$ represents the weight assigned to the global path, L_0 is the gate midpoint, and $\tau > 0$ governs transition smoothness. The default ZS-MTR-RR configuration employs fixed parameters $L_0 = 52$ and $\tau = 6$: shorter texts rely primarily on global discrepancy, whereas increasing scored length smoothly shifts weight to trajectory evidence.

This mechanism allocates evidence weights directly by scored length without training an auxiliary detector or router. Higher $S_{\mathrm{RR}}(x)$ indicates stronger evidence of machine generation.

4 Experiments

We address four research questions: **RQ1**, overall detection performance; **RQ2**, the contributions of the evidence paths and trajectory components; **RQ3**, the effectiveness of length-conditioned routing; and **RQ4**, detection performance under low false-positive rates.

4.1 Experimental Setup

Datasets. Our primary evaluation uses EvoBench Yu et al. (2025), covering XSum, PeerRead, PubMed, Writing, and Harmful across 51,873 common evaluation texts. The transformation evaluation contains 7,974 RealBench-derived texts He et al. (2025) spanning polishing, paraphrasing, perplexity attacks, and translation. The input budget is 512 tokens, providing at most 511 scored positions for ZS-MTR-RR.

Baselines. We compare statistical and recent zero-shot detectors, including Fast-DetectGPT Bao et al. (2024), Binoculars Hans et al. (2024), Lastde++ Xu et al. (2025), and Telescope Nassif & Cooper (2026), alongside supervised detectors such as RoBERTa Solaiman et al. (2019), MAGE Li et al. (2024), and RADAR Hu et al. (2023). Model configurations and inference details are provided in Appendix B. Component and routing controls reuse identical proxy outputs.

Metrics. AUROC is the primary metric. The primary EvoBench summary equally averages generator-family results within each domain, then equally averages the five domains. Transformation results are equally averaged over the four conditions. We additionally report PairAcc over complete human–AI pairs: the proportion in which the machine text receives a higher score, with ties counting as one half. Confidence intervals for key differences use 10,000 source-cluster bootstrap replicates, with multiple-comparison corrections detailed in Appendix C.

Evaluation Protocol. The main evaluation uses unlabeled target-pool medians and MADs for scale alignment; transfer results under fixed source-derived scales are reported separately. Method development for the original configuration utilized EvoBench results. To examine fusion choices, we additionally conduct retrospective source-disjoint five-fold diagnostics: within each fold, scales, static fusion weights, and gate parameters are estimated or selected using exclusively the development split and evaluated on the held-out split. Source-grouping rules, search grids, and selection criteria are detailed in Appendix C.

4.2 Overall Detection Performance (RQ1)

Table 1: Detection performance on 51,873 common EvoBench texts (%). Domain scores average generator families equally; Overall averages the five domains equally. Δ denotes percentage-point gains over Binoculars, computed from unrounded scores.

	Domain AUROC ↑					Overall ↑	
Method	**XSum**	**PeerRead**	**PubMed**	**Writing**	**Harmful**	**AUROC**	**PairAcc**
Supervised baselines							
RoBERTa-base	79.06	56.94	50.61	64.88	57.00	61.70	62.86
RoBERTa-large	76.16	60.76	55.11	55.95	61.67	61.93	64.15
MAGE	95.03	73.03	83.95	93.20	62.22	81.48	82.86
RADAR-Vicuna-7B	95.47	63.65	76.61	82.21	65.91	76.77	79.94
Zero-shot baselines							
Likelihood	70.76	72.31	79.44	87.41	46.09	71.20	74.08
Entropy	64.34	56.03	34.80	37.10	57.81	50.02	49.79
LogRank	72.31	73.23	78.21	87.24	45.37	71.27	74.22
DetectLLM-LRR	72.97	69.59	66.77	83.40	43.90	67.33	70.23
Fast-DetectGPT	96.81	83.73	89.48	98.36	57.50	85.18	87.15
Binoculars	96.94	83.96	88.82	98.37	57.86	85.19	87.28
DNA-DetectLLM	97.77	82.25	88.89	97.91	58.00	84.96	87.41
Lastde++	91.53	81.50	75.00	96.14	52.22	79.28	81.23
SpecDetect++	88.05	77.09	70.13	94.50	51.54	76.26	80.65
LAPD	85.60	61.37	84.32	91.18	59.83	76.46	79.86
Telescope	64.05	53.41	70.05	85.38	53.78	65.33	70.00
Uncertainty++	86.63	73.70	69.14	93.92	47.72	74.22	78.25
TempTest	90.41	82.08	82.18	96.45	52.41	80.71	82.70
ZS-MTR-RR (Ours)	95.22	90.87	91.61	95.52	64.41	**87.53**	**88.70**
Δ vs. Binoculars	-1.72	$+6.91$	$+2.80$	-2.85	$+6.56$	$\mathbf{+2.34}$	$\mathbf{+1.42}$

Table 1 summarizes detection performance across the five EvoBench domains. ZS-MTR-RR achieves a cross-domain average AUROC of **87.53%** and an average PairAcc of **88.70%**, outperforming Binoculars by **2.34** and **1.42 percentage points**, and Fast-DetectGPT by **2.35** and **1.55 points**, respectively. The 95% source-cluster bootstrap confidence interval for the AUROC gain over Binoculars is **[1.52, 3.15] percentage points**, substantiating the stability of the average lead under source-level resampling.

These improvements are primarily driven by PeerRead, PubMed, and Harmful, where ZS-MTR-RR exceeds Binoculars by **6.91**, **2.80**, and **6.56 AUROC points**, respectively. Meanwhile, on XSum and Writing, ZS-MTR-RR reaches **95.22%** and **95.52%**, though falling slightly behind that baseline. These results show that ZS-MTR-RR obtains a higher cross-domain average by securing substantial gains across challenging academic and safety domains while maintaining competitive accuracy on standard benchmarks.

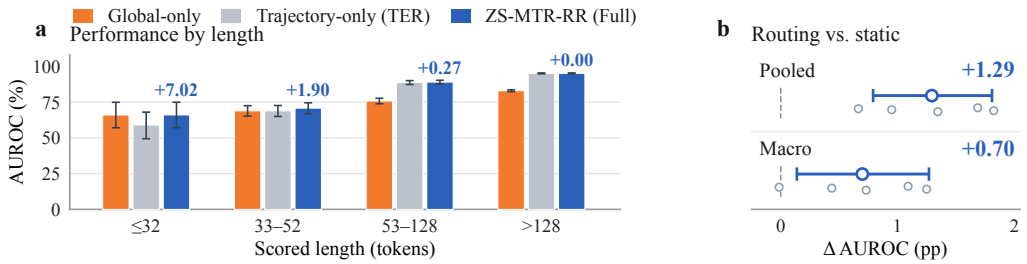

Figure 3: Length-conditioned routing analysis. (a) AUROC across four natural-length groups with 95% bootstrap CIs; annotations show gains over Trajectory-only. (b) Five-fold gains of development-selected routing over static fusion with Bonferroni CIs (10,000 replicates).

4.3 Contributions of Evidence Paths and Trajectory Components (RQ2)

Table 2: Ablation of evidence paths and trajectory components on EvoBench (AUROC, %). Domain columns and Avg follow Table 1's averaging convention. All variants share proxy outputs and scalers. Δ denotes percentage-point differences relative to the full method.

	Domain AUROC ↑					Overall ↑	
Method Variant	**XSum**	**PeerRead**	**PubMed**	**Writing**	**Harmful**	**Avg**	**Δ**
Component ablations							
w/o Stability V	87.54	75.73	88.29	89.20	64.65	81.08	−6.44
w/o Proxy tension A	90.14	90.57	86.55	94.75	62.10	84.82	−2.70
w/o Reuse group (H, G, E)	94.08	89.91	91.51	93.46	63.68	86.53	−1.00
Single-path baselines							
Trajectory-only (TER)	95.23	91.43	90.62	95.39	60.32	86.60	−0.93
Global-only (D_z)	83.38	67.00	85.47	84.49	63.46	76.76	−10.77
ZS-MTR-RR (Ours)	95.22	90.87	91.61	95.52	64.41	**87.53**	—

Primary Component and Path Contributions. Table 2 compares the complete method against component ablations and single-path variants. Full ZS-MTR-RR achieves an average AUROC of **87.53%**. Removing late-stage stability V or proxy tension A decreases average AUROC to **81.08%** and **84.82%**, representing drops of **6.44** and **2.70 percentage points**, confirming their primary discriminative role. The full framework also outperforms both single paths: retaining only the trajectory path or the global path yields average AUROCs of **86.60%** and **76.76%**, respectively. Combining both paths yields an overall gain of **0.93 points** over the trajectory-only model, with notable improvements in Harmful (**+4.09 points**) and PubMed (**+1.00 point**).

Value of the Full Representation. To examine whether the complete multi-trajectory representation is necessary beyond a simpler combination, we keep the global path and length gate unchanged and ablate the reuse feature group (H, G, E), reducing the trajectory path to $V + A$. This simplified variant achieves **86.53%** average AUROC, trailing the full method by **1.00 percentage point**. The full method achieves higher AUROC across all five domains, with a **2.06-point** gain on Writing. Under source-disjoint five-fold diagnostics, the full representation yields positive fold-wise gains over the $V + A$ baseline in every fold, with an average gain of **1.00 percentage point** and a multiple-comparison corrected bootstrap confidence interval of **[0.83, 1.19] percentage points**. These results confirm that the reuse group provides non-redundant discriminative signals beyond stability and tension. Individual ablations of auxiliary features are documented in Appendix D.

4.4 Effectiveness of Length-Conditioned Routing (RQ3)

Evidence Complementarity Across Lengths. Figure 3(a) compares the performance of the global path, trajectory path, and full ZS-MTR-RR across four natural-length groups. On passages of

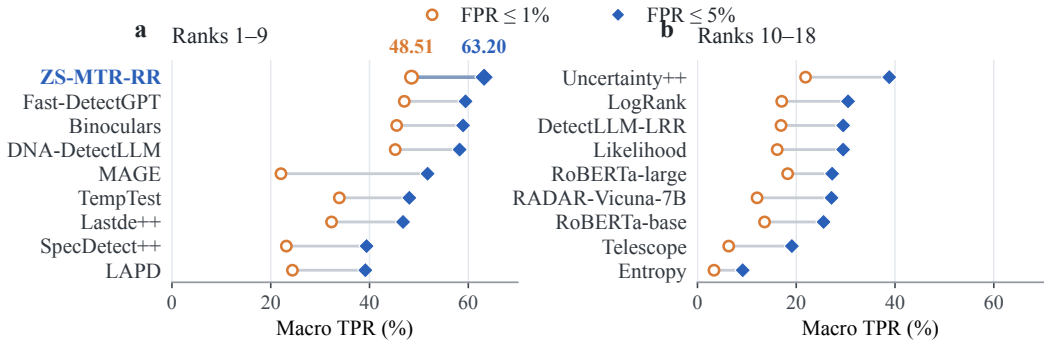

Figure 4: Low-FPR detection on EvoBench. Open circles and filled diamonds show macro TPR at empirical FPR $\leq 1\%$ and $\leq 5\%$, respectively, averaged over five domains. Methods are ordered by macro TPR at 5% FPR, continuing from (a) to (b).

length $L \leq 32$, the global path achieves higher AUROC than the trajectory path, whereas trajectory evidence establishes an increasing advantage as sequence length grows. Length-conditioned routing leverages this divergence, yielding gains of **7.02** and **1.90 percentage points** over the trajectory-only baseline in the first two length intervals; for sequences longer than 128 tokens, both the full method and the trajectory path reach **95.02%**. These results confirm that dynamic routing substantially enhances detection on short texts while preserving the discriminative capacity of trajectory statistics on longer documents.

Gains Over Static Fusion. To evaluate whether length-conditioned allocation surpasses fixed-weight combinations, we conduct retrospective source-disjoint five-fold diagnostics, selecting gate parameters and static fusion weights independently within each development fold and evaluating on held-out folds. As shown in Figure 3(b), length routing achieves an average fold-level pooled AUROC improvement of **1.29 percentage points**, with a multiple-comparison corrected bootstrap confidence interval of **[0.79, 1.81] percentage points**; macro-average AUROC improves by **0.70 percentage points** with a corresponding interval of **[0.14, 1.27] points**. Pooled gains remain strictly positive across all five held-out folds, confirming consistent benefits over development-selected static fusion. Further length-stratified decomposition indicates that overall gains stem primarily from improved cross-length ranking consistency, as detailed in Appendix D.

4.5 Low-FPR Detection Performance (RQ4)

Under domain-specific thresholding, ZS-MTR-RR achieves macro-average true positive rates (TPR) of **48.51%** and **63.20%** under 1% and 5% false positive rate (FPR) constraints, respectively, outperforming all 17 evaluated baselines (Figure 4). Compared to the strongest baseline at both thresholds, Fast-DetectGPT, this represents gains of **1.47** and **3.76 percentage points**, demonstrating that our average detection advantage extends into domain-specific low-FPR evaluation. This advantage depends on threshold protocol: under a global pooled threshold, Fast-DetectGPT achieves higher TPR at 1% FPR, with detailed threshold comparisons documented in Appendix E.

5 Conclusion

We presented ZS-MTR-RR, a training-free detector that combines global proxy discrepancy with position-sensitive trajectory evidence through a deterministic length-conditioned gate. Under the unified EvoBench protocol, it achieves the highest cross-domain average AUROC among the evaluated methods. Component ablations and retrospective source-disjoint diagnostics support the value of the complete trajectory representation and length-conditioned fusion. These findings show that scored length can serve as an effective, lightweight conditioning variable for coordinating complementary evidence without training an additional detector or learned router.

Reproducibility Statement. This study uses public benchmark datasets (EvoBench and four RealBench-derived transformations) and frozen open-weight reference models (Falcon-7B and Falcon-7B-Instruct), evaluating all methods under a unified protocol across all subsets and carriers without additional model training. The two evidence paths are computed on shared proxy outputs over the same valid scored positions, varying only the fusion weights with scored length. Theoretical analysis and gate formulation are given in Section 3; implementation details (tokenization, feature boundaries, scoring, and routing) in Appendix A; data, metrics, and reference-pool protocol in Appendix B; source-disjoint five-fold diagnostics in Appendix C; length and component ablations in Appendix D; low-FPR evaluation in Appendix E; reference-pool size and transfer boundaries in Appendix F; and reproducibility requirements with evidence-boundary notes in Appendix G. Anonymized code with score-file hashes 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*, 2024. URL https://openreview.net/forum?id=Bpcgcr8E8Z.

Souradip Chakraborty, Amrit Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. Position: On the possibilities of AI-generated text detection. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6093–6115. PMLR, 2024. URL https://proceedings.mlr.press/v235/chakraborty24a.html.

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, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.674.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 111–116, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3019.

Yikai Guo, Bin Wang, Xilai Fan, Wenjun Ke, and Haoran Luo. On the salience of low-probability tokens for AI-generated text detection: A multiscale uncertainty perspective, 2026. URL https://arxiv.org/abs/2606.02158.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17519–17537. PMLR, 2024.

Yongxin He, Shan Zhang, Yixuan Cao, Lei Ma, and Ping Luo. DETree: DEtecting human-AI collaborative texts via tree-structured hierarchical representation learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 146060–146099. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d6f8517fceeca1e2cd61721dff786c14-Abstract-Conference.html.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. RADAR: Robust AI-text detection via adversarial learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15077–15095. Curran Associates, Inc., 2023. doi: 10.52202/075280-0662. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/30e15e5941ae0cdab7ef58cc8d59a4ca-Abstract-Conference.html.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1808–1822, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.164. URL https://aclanthology.org/2020.acl-main.164/.

Tom Kempton, Stuart Burrell, and Connor J. Cheverall. TempTest: Local normalization distortion and the detection of machine-generated text. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1972–1980. PMLR, 2025.

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, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.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.

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.

Shixuan Ma and Quan Wang. Zero-shot detection of LLM-generated text using token cohesiveness. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17538–17553, 2024.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24950–24962. PMLR, 2023.

Christopher Nassif and Josh F. Cooper. Telescope: Improving zero shot detection of LLM generated content by measuring token repetition probability, 2026. URL https://arxiv.org/abs/2607.04061.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. Release strategies and the social impacts of language models, 2019. URL https://arxiv.org/abs/1908.09203.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12395–12412, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.827.

Ke Sun, Guangsheng Bao, Han Cui, and Yue Zhang. Minimizing mismatch risk: A prototype-based routing framework for zero-shot LLM-generated text detection, 2026a. URL https://arxiv.org/abs/2602.01240.

Ke Sun, Guangsheng Bao, Han Cui, and Yue Zhang. When AI settles down: Late-stage stability as a signature of AI-generated text detection, 2026b. URL https://arxiv.org/abs/2601.04833.

Alva West, Yixuan Weng, Minjun Zhu, Luodan Zhang, Zhen Lin, Guangsheng Bao, and Yue Zhang. AI-generated text is non-stationary: Detection via temporal tomography, 2025. URL https://arxiv.org/abs/2508.01754.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338, 2025. doi: 10.1162/coli_a_00549.

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, Huangsen Cao, Zhouhan Lin, Yu Zhao, and Fei Wu. Training-free LLM-generated text detection by mining token probability sequences. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vo4AHjowKi.

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, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.754.

Xiaowei Zhu, Yubing Ren, Fang Fang, Qingfeng Tan, Shi Wang, and Yanan Cao. DNA-DetectLLM: Unveiling AI-generated text via a DNA-inspired mutation-repair paradigm. In *Advances in Neural Information Processing Systems*, volume 38, pp. 1745–1768. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0287c393907259ae6a269fca5e3509cd-Abstract-Conference.html.

A Implementation Details

A.1 Tokenization and Shared Proxy Outputs

The primary proxies are Falcon-7B and Falcon-7B-Instruct, with the Falcon tokenizer and a shared vocabulary of 65,024 tokens. No special tokens are added. Inputs contain at most 512 tokenizer IDs. For a truncated document, the selected prefix is decoded and re-tokenized; the recovered IDs must equal the selected IDs. No additional text normalization is applied. In $x = (x_0, \ldots, x_L)$, x_0 supplies the initial context, so L counts scored positions rather than input IDs. The evaluated feature records satisfy $8 \leq L \leq 511$.

Both proxies receive identical IDs and remain frozen. Model parameters use FP16, while logits are converted to FP32 for log-softmax. We cache observed-token log-probabilities and the top-100

candidate IDs with their log-probabilities. The residual probability mass is also stored. Floating-point cache fields use FP32; feature computation uses FP64. The proxy lengths, candidate-array shapes, token alignment, vocabulary bounds, and finiteness of retained values are checked before scoring. Each proxy requires one forward pass, with its outputs shared by both evidence paths. This describes the extraction cost, not a claim that the complete Transformer computation is linear in sequence length.

A.2 Feature Boundaries and Numerical Conventions

The local-volatility configuration value is 20, corresponding to a radius of 10 and at most 21 positions per centered window. Windows are truncated at sequence boundaries without padding. Standard deviations use the population denominator. The stability statistic averages valid windows in the latter half of the scored sequence and negates the result. Windows with fewer than three positions are skipped. The evaluation requirement $L \geq 8$ is stricter than the underlying stability routine's minimum length.

For reuse probe $t \in \{1, \ldots, L-1\}$, the next-token distribution is conditioned on $x_{\leq t}$ and queried for token x_t. Thus, reuse probes are shifted by one position relative to observed-token likelihoods. If r_t^m is the probability mass outside the retained top-κ set, the estimator is

$$\widehat{q}_t^m = \begin{cases} p_m(x_t \mid x_{\leq t}), & x_t \in \mathcal{T}_{t,\kappa}^m, \\ \max(r_t^m, \varepsilon)/(|\mathcal{V}| - \kappa), & \text{otherwise}, \end{cases} \tag{S1}$$

where $\kappa = 100$, $|\mathcal{V}| = 65{,}024$, and $\varepsilon = 10^{-8}$. Residual mass is clipped to $[0, 1]$. All logarithms are natural. The reuse gap averages $\log \widehat{q}_t^B - \log \widehat{q}_t^I$ over the latter half of the $L-1$ probes, including positions without observed adjacent repeats.

Recurrence uses lags $1, \ldots, \min(32, L-1)$. When total lag-matching mass is at most 10^{-8}, the raw recurrence statistic is unavailable and replaced by zero before scale estimation and transformation. Other missing or non-finite required features are not imputed; such records receive no score. Zero replacement does not distinguish absent recurrence from a negative entropy close to zero. We retain this convention rather than changing the feature definition after evaluation.

A.3 Score Composition and Routing

Each feature is centered by its reference-pool median and divided by its unscaled median absolute deviation (MAD). No Gaussian consistency factor is used. A MAD below 10^{-8} is replaced by 1, rather than clamped to 10^{-8}. The coefficients are $w_V = w_A = 1$ and $w_H = w_G = w_E = 0.2$. There is no intercept or additional path-level normalization. The default gate has midpoint 52 and temperature 6. Its exponent $(L - 52)/6$ is clipped to $[-60, 60]$ for numerical stability. Scores are ranking statistics, not calibrated probabilities.

These features build on earlier likelihood-discrepancy, stability, and reuse signals Liu et al. (2025); Sun et al. (2026b); Nassif & Cooper (2026). Median/MAD alignment is a standard statistical operation. The proposed design concerns the organization of these signals into two paths and their length-conditioned fusion; we do not claim each constituent statistic as an independent new detection principle.

B Evaluation and Development Protocol

B.1 Datasets and Aggregation

The primary EvoBench comparison uses 51,873 common records: 25,931 complete human–AI pairs and 11 unpaired records. All records contribute to AUROC; only complete pairs contribute to PairAcc. The primary AUROC summary first averages generator-family results equally within each domain, then averages the five domains equally. Pooled AUROC instead uses all records in a single calculation and includes cross-domain and cross-family score comparisons. These summaries are not interchangeable.

PairAcc is the fraction of complete pairs for which the AI text receives a higher score, with ties contributing one half. Score directions are fixed by each method; we do not reverse directions using

Table S1: Principal recorded zero-shot baseline settings.

Method	Configuration
Fast-DetectGPT	Falcon-7B sampling proxy; Falcon-7B-Instruct scoring proxy.
Binoculars	Official BF16 configuration; maximum length 512; saved score direction.
DNA-DetectLLM Zhu et al. (2025)	FP32 inference; internal maximum length 1024 on common input prefixes.
Lastde++ / SpecDetect++	100 samples; embedding size 4; $\epsilon = 8$; $\tau' = 15$; seed 0.
LAPD Wu et al. (2026)	$\lambda = 0$; add=false; 10,000 Monte Carlo samples per token; refreshed runs use text-derived deterministic seeds.
Telescope	Released detector settings under the shared input budget.
TempTest	$T = 0.8$.
Uncertainty++	Tail 13%; Rényi order 1.6; local weight 0.1; 100 samples; seed 42; BF16.

evaluation labels. The supplementary transformation panel contains 1,984 polished, 1,996 paraphrased, 1,994 perplexity-attacked, and 2,000 translated texts, for a total of 7,974. Its summary is the unweighted average of the four condition-specific AUROCs. Human and AI records do not constitute verified source pairs in this panel, so PairAcc is not reported. This four-condition panel was selected exploratorily during the original development period, before supervised-baseline scoring. It is not a new, untouched external test or the complete RAID benchmark.

B.2 Reference Pools and Historical Development

The primary EvoBench result uses unlabeled statistics from the complete target pool. Although scaler estimation does not access labels, it accesses the evaluation distribution and is therefore transductive. Component-wise alignment can change the ranking of the combined score; it is not merely threshold calibration.

The original feature composition, coefficients, and gate configuration were developed with reference to EvoBench results. We do not claim that they were selected exclusively on an independent development set. The retrospective search below tests fusion alternatives under source isolation; it does not reconstruct the original development search or remove historical model-selection effects. In particular, its gate grid is not evidence that the default 52/6 setting or the coefficient 0.2 was originally selected by that grid.

The archived transformation result uses a seven-condition unlabeled reference pool containing 13,994 feature records, while evaluation is restricted to the four conditions above. We distinguish this setting from re-estimating scalers on the 7,974 four-condition records and from applying scalers fixed on the EvoBench development subset. The feature definitions and default gate remain unchanged in these comparisons.

B.3 Baselines and Configuration Scope

The complete inventory comprises 13 zero-shot detectors and four supervised baselines. Their names and results are listed in Tables S9 and S11. Comparisons use the specified common records rather than method-dependent evaluation subsets. Inputs follow the common 512-token budget; methods retain their own tokenizers and scoring interfaces. Principal recorded settings are listed in Table S1.

Binoculars, LAPD, and DNA-DetectLLM were rescored with checked checkpoint loading under PyTorch 2.6.0 and Transformers 4.51.3. Other baseline rows use archived scores with their recorded configurations. Consequently, unified metric recomputation does not mean that every detector was rerun. Sample IDs, labels, metadata, and score-file hashes are retained for verification. Detector inference and post-hoc metric aggregation are treated as separate stages.

B.4 Uncertainty and Information Boundaries

The main comparison with refreshed Binoculars and the natural-length intervals use 10,000 source-cluster bootstrap replicates and unadjusted 95% percentile intervals. The five-fold comparisons use the corrected intervals defined in Section C. The external transformation and proxy-sensitivity in-

Table S2: Held-out AUROC (%). Gate lists selected midpoint/temperature. Macro uses domain/family balancing; pooled uses all held-out records within that fold.

Fold	Test n	α	Gate	Static macro	Gate macro	Static pooled	Gate pooled
1	9,575	0.10	48/6	86.69	87.78	87.25	88.94
2	11,065	0.15	48/6	86.76	86.74	86.54	87.20
3	9,452	0.05	48/4	87.42	87.86	88.04	88.99
4	10,891	0.10	48/6	86.09	87.33	86.56	88.38
5	10,890	0.10	48/6	87.43	88.16	86.62	87.96

tervals are unadjusted 95% source-cluster percentile intervals. Historical permutation and matched-length results are reported as diagnostic point summaries; their variation across seeds or subsamples is not a source-bootstrap confidence interval.

Reference-pool-size intervals describe variation over 50 random unlabeled pools, not uncertainty over an independently sampled test population. Low-FPR comparisons report empirical operating-point estimates without confidence intervals or significance claims. All differences are computed before display rounding. Bootstrap intervals condition on the fixed scoring rules and reference scalers; they do not account for all historical feature or model selection.

C Retrospective Source-Disjoint Diagnostics

C.1 Source Grouping and Fold Construction

Records sharing a human–AI pair identifier, an original-text hash, or an input-prefix token-ID hash are joined into connected components. Transitive links are retained, so duplicate-associated records remain together even when their direct pair identifiers differ. This yields 8,614 source components over the 51,873 evaluated records. Each component is assigned to one of five folds by hashing its canonical identifier with seed 20260925 and taking the resulting integer modulo five. Fold construction does not use labels or detector performance. There is zero source-component overlap between development and held-out records in every fold.

C.2 Development-Only Scale Estimation and Selection

For each held-out fold, the other four folds supply feature medians, MADs, and the labels used for fusion-parameter selection. Static fusion is $(1-\alpha)S_{\mathrm{TER}} + \alpha D_z$, with $\alpha \in \{0, 0.05, \ldots, 1\}$. The gate search uses $L_0 \in \{32, 40, 48, 52, 56, 64\}$ and $\tau \in \{4, 6, 8, 12\}$. Both searches maximize the same development domain/family-balanced macro AUROC. Exact ties retain the first candidate in the listed ascending grid order. Parameters are not chosen using held-out scores. Feature coefficients are fixed throughout.

We also evaluate the original fixed 52/6 gate and a simplified $V + A$ trajectory under that same fixed gate. Removing features does not renormalize the remaining coefficients. Table S2 reports the selected configurations and held-out comparisons.

C.3 Five-Fold Aggregation and Intervals

Results are equal-weight averages of the five within-fold metrics. Differently calibrated scores are not concatenated into one pooled evaluation. For uncertainty estimation, source components are sampled with replacement within each held-out fold. Every record in a component receives the same resampling weight. Differences are computed within folds and then averaged, using 10,000 replicates. Scalers and selected parameters remain fixed during resampling; the bootstrap does not repeat parameter selection.

The comparison family contains three contrasts—fixed gate versus selected static fusion, selected gate versus selected static fusion, and full versus simplified representation under the fixed gate—for two metrics. Bonferroni-adjusted percentile intervals therefore use quantiles $0.025/6$ and $1 - 0.025/6$. Table S3 reports all six contrasts. Selected-gate pooled gains are positive in every fold;

Table S3: Five-fold mean AUROC differences and Bonferroni-adjusted intervals (percentage points).

Comparison	Macro gain	Interval	Pooled gain	Interval
Fixed gate – static	+0.72	[0.11, 1.36]	+1.41	[0.84, 2.00]
Selected gate – static	+0.70	[0.14, 1.27]	+1.29	[0.79, 1.81]
Full – simplified (fixed gate)	+1.00	[0.83, 1.19]	+1.47	[1.26, 1.68]

Table S4: Natural-length performance (AUROC, %). Each row uses all records in that length group.

Scored length	Human	AI	Global	TER	RR
$L \leq 32$	581	1013	65.94	58.95	65.98
33–52	4039	3948	68.84	68.83	70.73
53–128	10704	10248	75.71	88.65	88.92
$L > 128$	10616	10724	82.90	95.02	95.02

its macro gain is slightly negative in one fold. The full-versus-simplified gains are positive in every fold for both metrics. These are retrospective diagnostics on the development-era benchmark, not a new independent blind evaluation.

D Length and Component Analyses

D.1 Natural Length and Cross-Length Ranking

Natural-length groups are formed from the scored length of the retained text, without further truncating it. They differ from the prefix-budget experiment below. Table S4 reports the group sizes and three path scores. Figure 3 uses unadjusted 95% source-cluster bootstrap intervals on these groups. Components may contribute records to multiple length groups; bootstrap weights preserve their shared source membership.

For score function s, let n_b^A and n_c^H be group counts and N_A, N_H the corresponding class totals. To separate within-group and cross-group comparisons, let $A_{bc}(s)$ be AUROC comparing AI records in length group b against human records in group c. If n_b^A and n_c^H are their counts, pooled AUROC is

$$A(s) = \sum_{b,c} \frac{n_b^A n_c^H}{N_A N_H} A_{bc}(s). \tag{S2}$$

Applying this identity to RR minus TER gives a total gain of 1.5037 percentage points: 0.0954 from same-group comparisons and 1.4084 from cross-group comparisons. These are weighted contributions, not averages of group AUROCs. The result shows that most of the pooled improvement comes from changes in cross-length rankings under this partition. It does not establish that the gate selects an optimal path within every length group or directly estimates statistical reliability.

D.2 Extended Pooled Ablations

Table S5 supplements the main-paper macro-average ablation table with pooled summaries and individual feature removals. All variants use the same 51,873 records. The hard gate chooses the global path at $L \leq 52$ and TER otherwise. Equal fusion assigns weight 0.5 to each path; reversed routing replaces $g(L)$ by $1 - g(L)$. Single-feature removals retain all other weights, scalers, and the default gate. Pooled and macro-average differences answer different questions and should not be substituted for each other.

D.3 Same-Cache Controls for Related Signals

To distinguish the fusion framework from its constituent signals, we also evaluate related formulas using the same cached Falcon outputs. Table S6 includes the instruction-minus-base sequence log-likelihood sum, temporal stability discrepancy (TSD), and TSD added to Fast-DetectGPT. These

Group	Variant	AUROC	PairAcc	Δ AUROC (pt.)
Full method	ZS-MTR-RR	0.8838	0.8925	0.00
Routing	Hard gate	0.8805	0.8907	−0.33
	Equal average	0.8612	0.8835	−2.26
	Reversed gate	0.7751	0.8286	−10.88
Evidence path	Trajectory only	0.8688	0.8826	−1.50
	Global only	0.7662	0.8247	−11.77
Component	Without late-stage stability V	0.8058	0.8281	−7.80
	Without proxy tension A	0.8481	0.8598	−3.58
	Without immediate reuse H	0.8794	0.8893	−0.45
	Without reuse gap G	0.8760	0.8875	−0.78
	Without recurrence E	0.8822	0.8918	−0.17

Table S5: Pooled routing and component ablations on EvoBench. Δ AUROC is the change from the full method in percentage points.

Table S6: Same-cache formula controls on EvoBench (AUROC, %). Macro uses the main domain/family-balanced convention.

Score	Macro	Pooled
Sequence proxy log-ratio	76.97	73.82
TSD (base)	77.92	77.82
TSD (instruction)	77.35	77.17
TSD (base) + Fast-DetectGPT	84.98	85.16
TSD (instruction) + Fast-DetectGPT	85.27	85.28
ZS-MTR-RR	87.53	88.38

are matched-proxy formula controls, not claims of reproducing each prior work's native checkpoint and evaluation protocol. They supplement, rather than replace, the released baseline comparisons. Raw TOCSIN cohesiveness without its base-detector fusion is likewise not treated as the complete TOCSIN detector.

D.4 Position, Region, and Matched-Length Controls

Permutations synchronously rearrange cached token IDs and proxy fields; the models do not rescore the resulting text. Global discrepancy is unchanged, and the original scalers remain fixed. Reversal and half-swap also move records across the late-stage region. These controls therefore test indexed-position and region dependence, rather than isolating semantic word order. Region controls change the aggregation region. Exact-length matching balances human and AI counts at each scored length; the joint control additionally matches domain and generator family. Matched results average 100 subsamples. These controls use labels for diagnostic sampling, not detector fitting.

The length-only diagnostic has orientation-free AUROC 0.5107. Matching preserves a positive RR–TER difference, but does not establish that length-dependent score scaling is irrelevant or identify a unique causal explanation for routing gains.

D.5 Input-Prefix Budgets

Cached outputs are truncated to input budgets of 32, 64, 128, 256, and 512 tokens, after which features, target-pool scalers, and scores are recomputed. The default gate is unchanged. This alters the observations available for the same documents; it is not an experiment restricted to naturally short documents. Coverage is the fraction of scored token positions retained relative to the 512-token run, not measured runtime or memory savings. All rows contain 51,873 records.

Control	RR AUROC	TER AUROC	Δ RR (pt.)	Seeds
Temporal-order controls				
Original	0.8838	0.8688	0.00	1
Reverse	0.7098	0.6888	−17.40	1
Half swap	0.7369	0.7094	−14.69	1
Block shuffle 1	0.8069	0.7863	−7.70	10
Block shuffle 8	0.8101	0.7890	−7.37	10
Block shuffle 16	0.8101	0.7874	−7.38	10
Block shuffle 32	0.8066	0.7887	−7.73	10
Sequence-region controls				
Late half	0.8838	0.8688	0.00	1
Full sequence	0.8616	0.8525	−2.23	1
Early half	0.7417	0.7135	−14.21	1
Middle half	0.8468	0.8263	−3.71	1
Last quarter	0.8532	0.8343	−3.06	1

Length control	RR AUROC	TER AUROC	RR–TER (pt.)	Records
Original	0.8838	0.8688	1.50	51,873
Exact-length matched	0.8866	0.8731	1.35	43,416
Domain+family+exact-length matched	0.8851	0.8703	1.48	35,004

Table S7: Temporal and length controls. Temporal and region rows report pooled AUROC; block-shuffle values average ten seeds. Matched-length rows average 100 balanced subsamples.

Table S8: Prefix-budget sensitivity. AUROC is pooled (%); coverage is relative to the 512-token scored positions.

Input budget	Coverage (%)	RR	TER	Global
32	27.04	56.77	59.47	56.49
64	51.09	82.06	81.19	68.31
128	76.72	87.04	85.66	74.19
256	99.98	88.39	86.88	76.62
512	100.00	88.38	86.88	76.62

E LOW-FPR EVALUATION

For each method and domain, we enumerate thresholds on the saved score and select the maximum empirical TPR satisfying FPR $\leq q$, for $q \in \{0.01, 0.05\}$. Predictions use $s(x) \geq t$; equal-score groups cannot be split. The macro summary is the unweighted mean of five domain TPRs, without a generator-family averaging stage. This differs from the main AUROC summary. For the pooled control, a single threshold is used across all 51,873 records.

Both threshold protocols use evaluation labels to locate ROC operating points. They do not train a detector, but they also do not demonstrate false-positive control for an independently chosen deployment threshold. No score direction is selected or flipped using these labels. Table S9 reports all methods and both protocols. The values are point estimates, with no significance claim.

The highest macro TPR at both FPR bounds is obtained by ZS-MTR-RR. Its gains over Fast-DetectGPT are 1.47 and 3.76 percentage points. Under a single pooled threshold, Fast-DetectGPT is higher at 1% FPR (46.09% versus 40.55%), whereas RR is higher at 5% FPR. Thus, the low-FPR advantage depends on the threshold protocol and is not a universal operating-point dominance claim.

Table S9: TPR (%) at empirical FPR upper bounds of 1% and 5%. Macro uses a separate threshold per domain; pooled uses one threshold for the complete pool.

Method	Macro 1%	Macro 5%	Pooled 1%	Pooled 5%
ZS-MTR-RR	48.51	63.20	40.55	60.53
Fast-DetectGPT	47.04	59.44	46.09	58.83
Binoculars	45.50	58.94	36.16	58.98
DNA-DetectLLM	45.21	58.24	29.48	53.31
MAGE	22.07	51.76	22.64	50.40
TempTest	33.88	48.06	21.04	45.88
Lastde++	32.31	46.79	31.90	44.91
SpecDetect++	23.16	39.41	24.57	37.96
LAPD	24.39	39.16	24.01	36.68
Uncertainty++	21.89	38.84	21.05	35.51
LogRank	17.06	30.49	10.91	23.60
DetectLLM-LRR	16.92	29.51	12.18	21.76
Likelihood	16.15	29.49	9.58	22.02
RoBERTa-large	18.26	27.27	15.44	24.66
RADAR-Vicuna-7B	12.07	27.14	2.92	14.08
RoBERTa-base	13.58	25.53	15.10	23.05
Telescope	6.33	19.11	6.88	16.09
Entropy	3.35	9.16	1.49	7.26

Table S10: Reference-pool-size sensitivity: pooled AUROC (%) and empirical 95% intervals across 50 pools.

Pool size	RR mean	RR interval	TER mean
128	88.31	[87.75, 88.74]	86.80
512	88.32	[88.09, 88.57]	86.86
2,048	88.35	[88.20, 88.48]	86.87
8,192	88.38	[88.32, 88.44]	86.88
Full	88.38	—	86.88

F Reference Pools and Transfer Boundaries

F.1 Reference-Pool Size and Within-Benchmark Transfer

We sample 128, 512, 2,048, or 8,192 unlabeled EvoBench records without replacement, repeating each size 50 times. These records are drawn from the target evaluation pool; the experiment does not impose source isolation between the scaler pool and evaluated records. Reported intervals are empirical quantiles across sampled pools. The full-pool row is a single reference result, not an uncertainty interval.

Excluding each target domain from scaler estimation yields pooled RR AUROC 88.08%, compared with 88.38% using the full target pool. Excluding each target generator family yields 88.31%. These controls change the reference statistics within EvoBench. They are distinct from freezing an external source scaler and do not establish transfer to an unseen benchmark.

F.2 Text Transformations

Table S11 reports the four-condition comparison. RR uses the archived seven-condition reference pool specified in Section B.2. Baseline values are taken from the common-record evaluation, with refreshed Binoculars, LAPD, and DNA-DetectLLM scores covering all 7,974 records. This panel does not support the claim that RR has the highest external average AUROC.

F.3 Fixed-Source Versus Target-Pool Alignment

For the fixed-source setting, medians and MADs are estimated from the development portion of the source split corresponding to held-out fold 1 in Table S2. They are applied to all target records

Table S11: Four-condition transformation AUROC (%). Avg. is the unweighted condition mean; RR uses the original seven-condition reference pool.

Method	Polish	Paraphrase	PPL attack	Translate	Avg.
RoBERTa-base	63.47	67.90	44.07	84.36	64.95
RoBERTa-large	59.37	66.41	43.51	82.31	62.90
MAGE	73.47	76.75	90.20	87.20	81.90
RADAR-Vicuna-7B	77.64	79.20	86.07	88.81	82.93
Likelihood	62.46	64.26	78.04	76.15	70.23
Entropy	45.92	44.60	32.15	36.11	39.70
LogRank	63.38	65.58	78.89	78.27	71.53
DetectLLM-LRR	65.50	68.59	79.86	82.72	74.17
Fast-DetectGPT	77.83	74.42	82.06	86.98	80.32
Binoculars	80.22	79.45	81.16	87.25	82.02
Lastde++	73.80	72.46	76.97	88.10	77.83
SpecDetect++	70.97	69.17	73.80	89.11	75.76
LAPD	76.14	64.63	65.16	86.74	73.17
Telescope	62.39	56.37	58.54	67.37	61.17
Uncertainty++	74.04	74.29	79.58	92.00	79.98
TempTest	74.39	76.12	81.61	85.15	79.32
DNA-DetectLLM	81.07	80.51	83.41	87.83	83.21
ZS-MTR-RR	78.53	79.81	79.08	87.93	81.34

Table S12: Mean transformation AUROC (%) under different scale sources. All variants share the corresponding scaler and proxy outputs.

Reference pool	$V + A$ routing	TER	Static	Full RR
Target four-condition	79.07	81.11	81.22	80.93
Fixed source	79.74	81.68	81.89	81.69

without refitting. The static fusion weight is selected on that source development portion and equals 0.1. No target labels are used to select the static weight or gate. Table S12 keeps all 7,974 evaluation records fixed while changing the reference pool. Target alignment uses the four-condition pool, not the archived seven-condition pool.

Full RR exceeds simplified routing by 1.85 points with target scalers and 1.95 points with fixed-source scalers. Their unadjusted 95% source-cluster intervals are [1.46, 2.24] and [1.49, 2.42] points. External bootstrap sampling uses source-ID clusters stratified by human/AI class, without constructing pseudo-pairs. Source scalers and scoring parameters remain fixed over the 10,000 replicates.

Routing is below selected static fusion in both settings: differences are -0.30 points with interval $[-0.48,-0.13]$ and -0.20 points with interval $[-0.35,-0.05]$, respectively. With fixed-source scalers, RR and TER differ by only 0.005 points, with interval $[-0.14,0.15]$. Thus, transfer of the full representation's benefit does not imply transfer of an additional routing benefit.

F.4 Proxy-Configuration Sensitivity

Four configurations share 51,872 EvoBench records: 25,931 complete pairs and ten unpaired records. Each changes the proxy pair, tokenizer, inference precision where applicable, and target-pool feature scalers. Falcon and Llama use FP16; Qwen3 uses BF16 because FP16 produced non-finite logits. The 512-token budget, top-100 retention, feature coefficients, and 52/6 gate remain fixed. This is sensitivity to the complete configuration, not an isolated causal effect of model family.

Routing gains are positive for Falcon and Llama 3, small with an interval spanning zero for Qwen3, and negative for Llama 3.1. The primary Falcon length pattern is therefore not universal across proxy configurations.

Table S13: Domain/family-balanced macro AUROC (%) on the common proxy subset. Differences and unadjusted 95% source-cluster intervals are RR minus TER, in percentage points.

Configuration	TER	RR	Difference	Interval
Falcon-7B	86.86	87.53	+0.67	[0.18, 1.15]
Llama-3-8B	90.06	91.47	+1.41	[0.98, 1.84]
Qwen3-8B	85.52	85.62	+0.11	[-0.33, 0.56]
Llama-3.1-8B	89.17	87.70	-1.46	[-2.04, -0.87]

G Scope and Reproducibility Notes

The contribution is a simple, training-free organization and fusion of existing or adapted evidence signals. Length controls the amount and arrangement of observations entering these statistics; it is not a direct measurement of reliability. Overlapping windows and shared contexts create dependencies, so position counts should not be interpreted as independent sample sizes. The observed length patterns do not prove that the logistic gate is AUROC-optimal.

Primary results condition on the original method development and the chosen proxy pair. Source-disjoint retrospective diagnostics improve the transparency of fusion comparisons but do not undo that development history. Target-pool alignment is transductive; low-FPR ROC operating points do not provide a deployment guarantee. External and proxy comparisons show that ranking advantages and routing increments can weaken or reverse. Zero-filled recurrence also retains the ambiguity between missing matches and highly concentrated matching distributions.

Reproduction requires the same input prefixes and sample intersection, proxy checkpoints and tokenizers, numerical precision, feature definitions, scaler source, fixed score direction, and metric aggregation. Checkpoint-loading checks, input-ID matching, and saved score hashes are used to distinguish inference changes from aggregation changes. Auxiliary same-cache reconstructions or raw signals are not reported as complete reproductions of another detector unless its full scoring protocol is evaluated.