UNVEILING SPECTRAL MECHANISMS IN TRAINING-FREE LLM TEXT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid advancement of Large Language Models (LLMs) makes it increasingly difficult to distinguish human writing from machine-generated text. Training-free detection offers a scalable solution, yet common confidence-based metrics mainly measure average token probabilities and often miss the signal fluctuations that characterize human writing, which we call "generative vitality". Spectral analysis offers a way to capture this vitality, but its mechanism and practical boundaries remain underexplored. In this paper, we analyze spectral detection from both theoretical and empirical perspectives. We connect spectral energy to variance in proxy log-probability trajectories and explain how broader human token choices create the fluctuations used by frequency-domain indicators. We further show that the strength of this signal depends on text length and sampling range: spectral evidence is clearest for long, continuous, constrained generation, while short, fragmented, mixed, and edited settings require complementary confidence and fluctuation views. These findings clarify when frequency-domain detection works and provide guidance for future multi-dimensional detector design.

1 INTRODUCTION

The rapid advancement of Large Language Models (LLMs) has transformed digital content creation Crothers et al. (2023), while also blurring the distinction between texts written by humans and those generated by machines. This ambiguity creates risks such as misinformation Opdahl et al. (2023); Fang et al. (2024) and academic plagiarism Else (2023); Currie (2023). Among existing solutions, **training-free detection** Solaiman et al. (2019); Su et al. (2023); Gehrmann et al. (2019); Ippolito et al. (2019); Xu et al. (2024); Luo et al. (2025) is particularly attractive because it identifies LLM-generated text from statistical patterns such as probability curvature Bao et al. (2023); Mitchell et al. (2023) or entropy Ippolito et al. (2019), without requiring large supervised training sets.

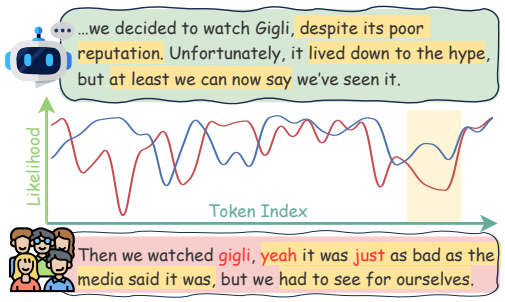


Figure 1: **Illustration of Generative Vitality.** The plot shows token log-likelihood trajectories computed by a proxy LLM. Human writing (red) contains vitality spikes caused by unpredictable low-probability tokens, marked in red in the text below. AI outputs (blue) are typically smoother because decoding favors high-probability tokens. Yellow-highlighted regions mark segments with large generation-probability divergence.

Most training-free methods rely on **confidence-based indicators** such as LogLikelihood and Rank Solaiman et al. (2019), which assume that LLMs select tokens with higher probabilities than humans. Recent studies Xu et al. (2024); Luo et al. (2025) suggest that these metrics overlook signal **fluctuation**, termed **generative vitality**. We define this vitality as the intermittent emergence of **unpredictable tokens** from a low-probability region. As visualized in Figure 1, human authors often introduce expressive low-probability tokens that create sharp negative spikes in the likelihood trajectory, while LLM decoding more often follows a smoother high-probability path.

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.

To capture these fluctuations, initial efforts such as Lastde Xu et al. (2024) use multi-scale entropy, but this can introduce sensitivity to hyperparameters. SpecDetect Luo et al. (2025) takes a frequency-domain approach, based on the observation that human writing tends to exhibit higher **spectral energy** than LLM-generated text. Although SpecDetect performs well, the mechanism behind this spectral gap and its behavior in complex real-world scenarios, such as short snippets, mixed-source content Lee et al. (2022); Wang et al. (2024), and collaborative editing Zhang et al. (2024), remain insufficiently understood. Consequently, the practical utility of frequency-domain detection is constrained by three critical "unknowns" that we address in this paper.

The Theoretical Unknown: *Why are frequency-domain methods effective?* Empirical results show a spectral energy gap, but the generation source of this gap remains unclear. **The Capability Unknown:** *What are the boundaries of their effectiveness?* Existing evaluations are primarily document-level, leaving short text, mixed-source text, and human-AI editing underexplored. **The Solution Unknown:** *How should we address spectral failure mode?* If spectral methods cannot excel in all scenarios, what is the solution? Specifically, can we integrate them with traditional methods to build a robust system?

We address these gaps through theoretical and empirical analysis. First, we model the generative vitality of human writing and show how fluctuation disparities manifest as a frequency-domain signature. We then evaluate spectral evidence across standard document-level benchmarks and challenging real-world scenarios, including mixed-source and collaborative writing.

Our analysis leads to several key insights. **First**, confidence and fluctuation provide complementary views of the same proxy probability trajectory. **Second**, their relative effectiveness changes with sequence length, sampling scope, source continuity, and editing operation, revealing the conditions under which each signal remains informative. Spectral evidence is clearest for long, continuous, constrained generation, while short fragments and point-wise editing favor confidence cues. **Finally**, adaptive weighting improves selected collaborative settings, while its transfer performance varies across operations. Together, these findings explain what spectral detectors measure and map the practical boundaries of confidence and fluctuation evidence.

Mapping where frequency-domain detection works and which signals inform different regimes provides a roadmap for multi-dimensional AI detectors. Our code is available in this repository.

2 Related Work

We provide a concise overview here and include a detailed discussion in Appendix A.

Training-Free LLM Text Detection. Training-free detection uses statistical signatures to identify machine-generated text. Distribution-based methods compare inputs with perturbed or generated variants Mitchell et al. (2023); Su et al. (2023); Yang et al. (2023); Bao et al. (2023), while sample-based methods directly score token-level statistics Solaiman et al. (2019); Su et al. (2023); Gehrmann et al. (2019); Ippolito et al. (2019); Xu et al. (2024); Luo et al. (2025). Sample-based indicators expose the probability signal directly. We study how frequency-domain detection models its fluctuation, using SpecDetect Luo et al. (2025) as a representative spectral method.

Detection in Complex Scenarios. Real-world content often contains mixed-source text Lee et al. (2022); Wang et al. (2024) and collaborative authorship Zhang et al. (2024), where human and machine segments are interleaved or edited together. Recent studies address these settings with supervised sequence-labeling or boundary-detection models Zeng et al. (2024); Su et al. (2025); Kadiyala et al. (2025), while training-free detectors are still mainly evaluated on long-form, single-source documents. We examine the utility and limitations of frequency-domain indicators when signals are fragmented by mixing or reshaped by collaborative editing.

3 Detection Signals and Spectral Mechanism

In the training-free LLM detection paradigm, text classification is transformed into mining patterns within the token probability sequence. As illustrated in Figure 2, a proxy LLM first converts text into token log-probabilities, and detectors extract features from this signal.

3.1 From Text to Probability Signals

Let $X = \{w_1, w_2, \ldots, w_N\}$ be a sequence of N tokens. To analyze the generation traces, existing works utilize a proxy LLM $\mathcal{M}$ (e.g., GPT-J Wang & Komatsuzaki (2021)) to compute the step-wise conditional probabilities. We define the probability signal as the log-probability of the token:

$$x[t] = \log P_{\mathcal{M}}(w_t \mid w_{<t}) \tag{1}$$

which forms the probability signal analyzed by training-free detectors. Given this signal, a detection algorithm computes an indicator $y = f(x)$, where larger y indicates stronger evidence of machine origin. An effective detector ranks machine text above human text in expectation, i.e., $\mathbb{E}[f(x_{\text{LLM}})] > \mathbb{E}[f(x_{\text{Human}})]$.

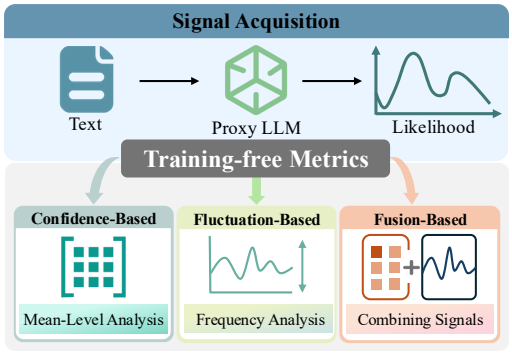


Figure 2: Overview of the Training-Free Detection Pipeline.

3.2 Metric Families: Confidence and Fluctuation

To evaluate how different methods extract information from $x[t]$, we categorize prevalent training-free metrics into a unified taxonomy in Table 1. We distinguish confidence-oriented indicators from fluctuation-oriented indicators; detailed formulations appear in Appendix C.1, and empirical relationships are examined in Section 4.2.

Confidence-based Metrics. Most existing methods focus on the global elevation of log-probabilities. Indicators such as LogLikelihood, LogRank Solaiman et al. (2019), LRR Su et al. (2023), and Entropy Gehrmann et al. (2019) capture the mean-shift signature, but are often insensitive to the structural rhythm of the text.

Fluctuation-based Metrics. Frequency-domain methods such as SpecDetect Luo et al. (2025) measure signal fluctuation. As derived below, LLM decoding suppresses human-like surprisal spikes, creating a detectable spectral signature.

Fusion-based Metrics. Hybrid methods combine confidence and fluctuation cues. Lastde Xu et al. (2024) combines likelihood with multi-scale entropy. We include SpecFusion as a probing indicator: it standardizes and sums SpecDetect and LogLikelihood to test whether mean-level and fluctuation evidence are complementary.

Table 1: **Taxonomy of training-free metrics.** We classify metrics by their statistical focus using checkmarks (✓) and cross marks (×). **Conf.** indicates confidence-based metrics, while **Fluct.** indicates fluctuation-based metrics.

Metric	Statistical Interpretation	Conf.	Fluct.
LogLikelihood (LL)	Measures average token log-probability.	✓	×
LogRank (LR)	Average of log-rank scores for token positions.	✓	×
LRR	Ratio of log-likelihood to log-rank (relative confidence).	✓	×
Entropy	Average predictive uncertainty of the distribution.	✓	×
SpecDetect	Frequency-domain modeling of fluctuations.	×	✓
Lastde	Ratio of log-likelihood to multi-scale entropy.	✓	✓
SpecFusion	Combination of LogLikelihood and SpecDetect.	✓	✓

3.3 Modeling Generative Vitality

We define generative vitality Luo et al. (2025) as intermittent visits to the scoring model's low-probability tail. Let $x[t] = \log P_{\mathcal{M}}(w_t \mid w_{<t})$ be the log-probability assigned by scoring model $\mathcal{M}$ at position t. At each position, $\mathcal{M}$ induces a context-dependent distribution over vocabulary $\mathcal{V}$. For a Top-p threshold τ, the **Head Set** $\mathcal{V}^{(t)}_{\text{head}}(\tau)$ is the smallest probability-sorted prefix whose cumulative mass reaches τ, while the **Tail Set** $\mathcal{V}^{(t)}_{\text{tail}}(\tau)$ contains the remaining lower-probability,

high-surprisal tokens. Top-k and related truncated decoding rules induce the same type of partition: they differ only in how the cutoff is selected, while our analysis uses the resulting high-probability region versus its complement. This local partition defines a signal-space boundary $\mathcal{L}^{(t)}_{\text{bound}} = \min_{w \in \mathcal{V}^{(t)}_{\text{head}}(\tau)} \log P_{\mathcal{M}}(w \mid w_{<t})$.

A Unified Mixture View. The head/tail event can be written with a single source-level mixture. Let $s \in \{H, \text{AI}\}$ denote the text source. For each token, define

$$x_s[t] = (1 - b^{(s)}_t) x^{(s)}_{\text{head}}[t] + b^{(s)}_t x^{(s)}_{\text{tail}}[t], \tag{2}$$

where $b^{(s)}_t = \mathbb{I}\{w_t \in \mathcal{V}^{(t)}_{\text{tail}}(\tau)\}$ indicates whether the token falls in the scoring model's tail set. The tail-token rate is $\gamma_s = \frac{1}{N}\sum_{t=1}^{N} b^{(s)}_t$, the fraction of positions where text source s enters the tail set.

Human and LLM text differ in this tail-token rate: human authors may choose locally unexpected tokens for specificity or style, whereas LLM decoding favors high-probability continuations. In a white-box setting where the scoring and source models are the same, truncated decoding keeps generated tokens mostly inside the head set; when the proxy and source models differ, some generated tokens may enter the proxy-defined tail. Since LLMs still tend toward high-probability tokens, sample-based detection methods commonly rely on the condition $\gamma_H > \gamma_{\text{AI}}$.

We examine this condition on the three standard datasets in Section 4.1, using GPT-J-6B and Llama3-8B as proxy scoring models. We measure two tail-entry rates: the fraction of tokens outside the Top-$p = 0.90$ head set and the fraction whose proxy rank is greater than 50. Figure 3 shows the GPT-J-6B Tail@0.90 result, and Appendix D.5 reports GPT-J-6B Rank>50 and the Llama3-8B counterparts in Figures 14, 15, and 16. Even under black-box proxy–source mismatch, human text enters the proxy tail region more often than AI text.

The Fluctuation Divergence. Holding the head/tail conditional means and variances fixed, we can isolate how tail-entry frequency contributes to the fluctuation gap.

Theorem 1 (The Fluctuation Divergence). *Let $\sigma^2_H = \text{Var}(x_H[t])$ and $\sigma^2_{\text{AI}} = \text{Var}(x_{\text{AI}}[t])$. The human–AI fluctuation divergence is $\Delta\sigma^2 = \sigma^2_H - \sigma^2_{\text{AI}}$. Under the unified mixture view, with shared head/tail conditional statistics across text sources,*

$$\begin{aligned}\Delta\sigma^2 = (\gamma_H - \gamma_{\text{AI}})(\sigma^2_{\text{tail}} - \sigma^2_{\text{head}}) \\ + \big[\gamma_H(1-\gamma_H) \\ - \gamma_{\text{AI}}(1-\gamma_{\text{AI}})\big](\mu_{\text{tail}} - \mu_{\text{head}})^2.\end{aligned} \tag{3}$$

The proof is in Appendix B.1. The formula explains the expected higher human fluctuation under the tail-rate condition. The first term is positive when human text enters the tail more often and tail-region log-probabilities are more dispersed than head-region values. The second term reflects mean separation: head tokens have log-probabilities close to zero, whereas tail tokens have much lower values. In the usual sparse-tail regime, $\gamma_H > \gamma_{\text{AI}}$ also makes this term positive, so both terms increase the human–AI variance gap.

The positive human–LLM variance gap also appears empirically under GPT-J-6B scoring when each source has its own head/tail conditional statistics (Appendix D.5.1).

3.4 Spectral Energy as a Fluctuation Signature

The variance gap in Theorem 1 transfers to spectral energy through Parseval's identity.

Corollary 2 (Spectral Energy Gap). *For a centered signal $z_s[t] = x_s[t] - \mu_s$, let $\bar{\mathcal{E}}_s = \frac{1}{N}\sum_{k=0}^{N-1} |\mathcal{F}(z_s)_k|^2$ be its average spectral energy. Under the same mixture view,*

$$\Delta\bar{\mathcal{E}} = \mathbb{E}[\bar{\mathcal{E}}_H] - \mathbb{E}[\bar{\mathcal{E}}_{\text{AI}}] = N\Delta\sigma^2. \tag{4}$$

The proof is in Appendix B.2. For the centered probability signal, Parseval's identity Hardy & Titchmarsh (1931) maps the larger time-domain variance of human text to larger expected spectral energy. LLM decoding suppresses these fluctuations, yielding lower expected spectral energy for machine-generated text. SpecDetect therefore uses negative energy as the AI-likeness score so larger values indicate stronger machine evidence.

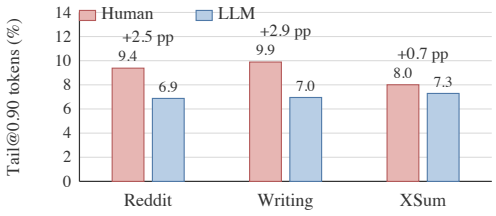


Figure 3: **Preliminary head/tail check.** Bars show the Tail@0.90 rate under GPT-J-6B proxy scoring. Human continuations generally use more tokens outside the Top-$p = 0.90$ head window than paired LLM continuations, especially on WritingPrompts and Reddit. The gap is smaller on XSum.

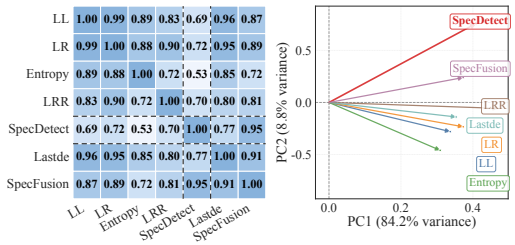


Figure 4: **Metric Relations on Writing Dataset.** (Left) **Correlation Heatmap**: Metrics cluster into Confidence and Fluctuation families. (Right) **PCA Loading Plot**: PC1 (84.2%) and PC2 (8.8%) capture the dominant variance directions.

Impact of Sequence Length. Sequence length N controls how reliably spectral evidence can be observed. Short texts may contain too few tail events or vitality spikes, while longer texts give the variance gap more opportunities to appear in the probability trajectory. This motivates the length sensitivity study in RQ2.

Impact of Sampling Scope. Sampling scope affects $\Delta\sigma^2 = \sigma_H^2 - \sigma_{\text{AI}}^2$. Restricted decoding keeps machine text in high-probability regions, whereas larger Top-p, Top-k, or temperature lets generated text enter the tail more often, raising σ_{AI}^2 and narrowing the spectral gap. This motivates the sampling-scope analysis in RQ3.

4 Empirical Evaluation in Standard Scenarios

We evaluate the spectral mechanism through three questions. **RQ 1** asks how frequency-domain indicators relate to existing methods and whether they capture a **distinct analytical dimension**; **RQ 2** examines how **sequence length** affects spectral methods relative to other indicators in short and long texts; and **RQ 3** tests their performance when stochastic sampling expands the **fluctuation scope** of machine text through higher Temperature, Top-k, or Top-p.

4.1 Experimental Setup

Our setup follows established document-level zero-shot protocols Xu et al. (2024); Luo et al. (2025). We evaluate three English datasets: **XSum** Narayan et al. (2018) for BBC news summarization, **WritingPrompts** Fan et al. (2018) for creative story generation, and **Reddit ELI5** Fan et al. (2019) for explanatory question answering; WMT16 Bojar et al. (2016) is in Appendix D.7 for language transfer checks. Each English dataset contains 150 human-written examples. We use the first 30 tokens of each human text as a prompt and generate a paired machine continuation, yielding balanced 150/150 human-machine evaluation sets with shared prompts and matched lengths.

Following previous work Xu et al. (2024); Luo et al. (2025), we evaluate all methods under a realistic black-box setting where the detector has no access to the source-model internals or identity. The main experiments use Llama2-13B Touvron et al. (2023) as the source model and GPT-J-6B Wang & Komatsuzaki (2021) as the proxy model for extracting $x[t]$. Additional source-model details are provided in Appendix C.3. Additional proxy-model results are reported in Appendix D.6. Unless specified otherwise, generation uses Temperature $T = 1.0$, Top-$p = 1.0$, Top-$k = 50$, and 150-word continuations. Following prior work Xu et al. (2024); Luo et al. (2025), we rank human and machine samples by detector score and report Area Under the ROC Curve (AUC) Faraggi & Reiser (2002).

We assess sampling uncertainty using 2,000 bootstrap resamples; the protocol and representative comparisons are reported in Appendix E.

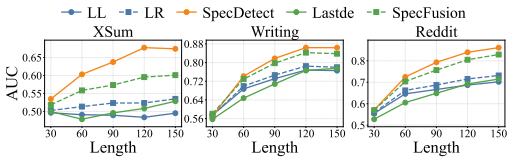


Figure 5: **Performance vs. Sequence Length on Three Datasets.** Fluctuation-based metrics achieve more significant gains as sequence length increases.

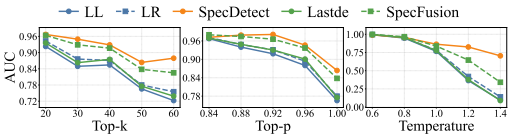


Figure 6: **Detection Robustness vs. Sampling Parameters on XSum.** Performance decays across all methods as Top-k, Top-p, and Temperature increase, yet fluctuation-based indicators demonstrate superior resilience compared to confidence-based metrics.

4.2 RQ 1: Validation of Metric Distinctness

For RQ1, we compute Spearman correlations Wissler (1905) and Principal Component Analysis (PCA) Abdi & Williams (2010) over detector scores. Figure 4 (Metric Relations on Writing Dataset) shows Writing results; other datasets appear in Appendix D.1, and input-granularity preferences are discussed in Appendix D.2.

Distinct Detection Dimension Figure 4 shows separation between confidence and fluctuation indicators. In the heatmap, LogLikelihood, LogRank, and Entropy cluster together, whereas SpecDetect has lower correlation with this confidence group. The PCA view confirms this: PC1 (84.2%) captures the common detection signal, while PC2 (8.8%) separates SpecDetect from mean-probability metrics. This supports the claim that spectral methods capture a fluctuation signature distinct from average probability level.

4.3 RQ2: Sensitivity to Sequence Length

Guided by Section 3, RQ2 truncates test samples to $L \in \{30, 60, 90, 120, 150\}$ words and tracks detector AUC in Figure 5. The figure highlights representative metrics. Complete main-setting length curves are reported in Appendix D.3.1. Additional source-model length results are reported in Appendix D.4.1. The Llama3-8B proxy counterpart is reported in Appendix D.6.1. The cross-language length check is reported in Appendix D.7.

Length Dependent Scaling Figure 5 shows that spectral detection benefits from longer context. At $L = 30$, SpecDetect is weak because short samples may contain too few tail events for a stable fluctuation estimate. As length grows, these local probability changes become easier to observe, and fluctuation-based indicators gain more than confidence-based metrics; fusion indicators typically fall between the two families. Paired-bootstrap comparisons support the stronger length-related gains of SpecDetect relative to LogLikelihood across the three datasets (Appendix E).

4.4 RQ3: Sensitivity to Sampling Scope

For RQ3, we vary stochastic decoding parameters around the default setting (Top-$k = 50$, Top-$p = 1.0$, Temperature $T = 1.0$): Top-$k \in \{20, \ldots, 60\}$, Top-$p \in \{0.84, \ldots, 1.0\}$, and $T \in \{0.6, \ldots, 1.4\}$. Figure 6 highlights representative metrics. Appendices D.3.2, D.4.2, D.6.2, and D.7 provide complete main-setting curves, additional source results, Llama3-8B proxy results, and cross-language checks, respectively.

Stochastic Decoding as a Bottleneck Figure 6 shows a decline as sampling becomes broader. Increasing k, p, or T lets generated text access the tail region more often, narrowing the human–machine fluctuation gap and reducing AUC across methods. At extreme temperatures ($T > 1.2$), machine text can become more volatile than human text, producing performance inversion; this conditional observation is consistent with the variance-energy mechanism in Corollary 2.

Spectral Resilience vs. Confidence Degradation All indicators suffer under high entropy, yet fluctuation-based metrics degrade more slowly than confidence-based metrics in this setting. As

sampling randomness increases, LogLikelihood becomes less distinguishable from human writing, while spectral signatures retain part of the fine-grained variance signal. Fusion indicators generally sit between the two families because they aggregate confidence and fluctuation views. Bootstrap analyses further support the sampling-scope findings (Appendix E).

4.5 Concluding Takeaways

The standard-scenario evaluation yields three takeaways. **Distinct Statistical Properties**: frequency-domain indicators capture fluctuation signatures that are distinct from confidence-based mean-probability metrics. **Boundary Conditions**: spectral detection benefits from sufficient sequence length, but degrades as stochastic decoding increases machine variance toward, or beyond, human levels. **Metric Complementarity**: confidence and fluctuation metrics provide safety nets across regimes; spectral structure remains useful when confidence signals flatten under high-entropy decoding, while confidence helps when short texts lack context for stable fluctuation estimates.

5 Empirical Evaluation in the Wild

Following the document-level analysis, we evaluate spectral methods in real-world regimes where human and AI-authored sentences can be interleaved Wang et al. (2024); Lee et al. (2022), edited, or humanized Zhang et al. (2024). We ask two questions: **RQ1 (Mixed Source Text)** studies how structural mixing affects spectral detection when the signal is fragmented, and **RQ2 (Collaborative Text)** studies how specific editing operations reshape the fluctuations used for detection. Our goal is to identify when frequency-domain evidence remains useful and when confidence-based cues become a necessary complement.

5.1 Experimental Setup

Scenarios and Datasets. Dataset statistics are provided in Appendix C.2. **Scenario I: Mixed-Source Text** covers documents where human-authored and AI-generated sentences coexist: **SemEval** Wang et al. (2024) uses a human prefix followed by a GPT-3 continuation with variable transition points and ratios, while **CoAuthor** Lee et al. (2022) interleaves Human, LLM, and Human-LLM Collaborative sentences throughout a document.

Scenario II: Collaborative Text examines how editing operations transform a text's statistical signature. CoAuthor provides a sentence-level view through its Human-LLM Collaborative category, while **MixText** Zhang et al. (2024) provides an operation-level benchmark with 300 original samples per direction. MixText includes **AI-Polishing** (Human $\rightarrow$ AI), covering token/sentence *Polish*, *Rewrite*, and *Complete*; and **Humanizing** (AI $\rightarrow$ Human), covering token/sentence *Humanize* and *Adapt*, generated by GPT-4 and Llama-2-70b.

Implementation and Evaluation. We use GPT-J-6B Wang & Komatsuzaki (2021) for all metrics, following the document-level settings. Scenario I reports average within-document AUC from sentence-level scores. Scenario II uses MixText pairwise accuracy: the percentage of pairs with the correct score shift after editing, such as AI-polished text scoring higher for machine origin than its human original. Appendix C.3 details the protocol.

5.2 RQ 1: Performance in Mixed Source Text

SemEval focuses on Human (H) vs. LLM (L) segments, while CoAuthor adds a Collaborative (C) category and supports H-L, H-C, and L-C comparisons. Table 2 reports the results, and the Llama3-8B proxy counterpart is in Appendix D.6.3. Lastde is omitted from this sentence-level table because its multi-scale diversity entropy component requires longer sequences than sentences provide.

Confidence Superiority in Fragmented Contexts Sentence-level units give spectral methods short observation windows, so pure-generation rows favor confidence metrics: LogLikelihood outperforms SpecDetect on SemEval (0.9085 vs. 0.8411), and LogRank leads on CoAuthor H vs. L (0.8138). This matches length sensitivity in Section 4.3: confidence metrics exploit mean-probability shifts, while fluctuation indicators need longer spans to accumulate variance evidence.

Blending Effect in Collaborative Detection Collaborative rows blend both sources' signatures: in CoAuthor, LogLikelihood drops from 0.8044 on H vs. L to 0.6459 on H vs. C, and LogRank falls from 0.8138 to 0.6514. This blending weakens the low-perplexity footprint of machine text while disrupting human fluctuation patterns.

Stability vs. Optimality in Fusion SpecFusion combines confidence and fluctuation evidence, but its gains depend on the operation: averaging the two channels can also dilute the more informative signal. In H-vs-C, its AUC is 0.6806, while SpecDetect reaches 0.7017. Document/session bootstrap comparisons further support the contrast between confidence-favored pure-source settings and fluctuation-favored H-vs-C detection (Appendix E).

Table 2: **Sentence-level Detection across Pure and Collaborative Text.** Average within-document AUC for binary classification among Human (H), LLM (L), and Collaborative (C) sentences. Best results are **bolded**; second-best are underlined.

Metric	Pure Generation		Collaborative Mixed	
	SemEval	CoAuthor (H vs. L)	CoAuthor (H vs. C)	CoAuthor (L vs. C)
Confidence-based				
LogLikelihood	**0.9085**	0.8044	0.6459	0.6951
LogRank	0.9038	**0.8138**	0.6514	**0.7010**
LRR	0.2544	0.3698	0.4425	0.4081
Entropy	0.8655	0.4322	0.4701	0.4759
Fluctuation-based				
SpecDetect	0.8411	0.7532	**0.7017**	0.5512
Fusion Indicator				
SpecFusion	0.8998	0.7988	0.6806	0.6497

5.3 RQ 2: Performance in Collaborative Text

Beyond structural splicing, MixText lets us dissect how collaborative operations transform scores. We use pairwise accuracy, measuring whether a detector assigns a higher machine-origin score to the version with stronger machine involvement. Table 3 reports results, Figure 7 visualizes AI-Polishing signal changes, and the Humanizing visualization is in Appendix D.8. The Llama3-8B proxy version is reported in Appendix D.6.4.

A further edit-density diagnostic shows that surface edit amount alone has limited explanatory power for MixText behavior; operation type, edit direction, and continuity are more informative. Full definitions and task-level summaries are provided in Appendix D.8.

5.3.1 AI-Polishing: From Local Repairs to Global Recasting

Table 3 and Figure 7 reveal regimes governed by the continuity of LLM involvement.

Spectral Blindness in Point-wise Polishing Sparse token- or sentence-level polishing raises local confidence while leaving the underlying human fluctuation mostly intact. Confidence metrics capture this uplift: in GPT-4 Polish Token, Entropy reaches 78.67% and LogLikelihood reaches 65.00%, while SpecDetect remains near chance at 52.67%. This source-inertia effect occurs because selected high-probability replacements rarely recast the sequence rhythm, so spectral signatures stay close to the human source. Paired-bootstrap differences support the reversal from LogLikelihood-favored GPT-4 token polishing to SpecDetect-favored completion (Appendix E).

Table 3: **Pairwise Accuracy on MixText Scenarios.** Values are the percentage of pairs where the Modified version is ranked more machine-like than the Original. Columns are ordered by machine involvement. Best results are **bolded**; second-best are underlined.

Metric	AI-Polishing (GPT-4)				AI-Polishing (Llama-2)				Humanizing			
	Polish Token	Polish Sentence	Complete (1/3+2/3)	Rewrite (Re-gen)	Polish Token	Polish Sentence	Complete (1/3+2/3)	Rewrite (Re-gen)	Adapt Token	Adapt Sentence	Llama-2 Humanize	GPT-4 Humanize
Confidence-based Metrics												
LogLikelihood	0.6500	0.6167	0.6500	0.4067	0.7433	0.8233	0.8100	0.9533	0.7267	0.8300	**0.7333**	**0.9667**
LogRank	0.6033	0.5800	0.6667	0.4233	0.6967	**0.8300**	0.7967	**0.9600**	0.7133	0.8300	**0.7333**	0.9600
LRR	0.3067	0.3133	0.5300	**0.4933**	0.3067	0.2700	0.1733	0.2767	0.2100	0.2167	0.2900	0.0800
Entropy	**0.7867**	**0.7067**	0.5067	0.4867	**0.8233**	0.8167	0.8300	0.8300	**0.7733**	0.7600	0.6967	0.9600
Fluctuation-based Metrics												
SpecDetect	0.5267	0.5967	**0.8100**	0.4733	0.6400	0.7967	**0.9700**	0.7967	0.6767	0.7700	0.7067	0.8833
Hybrid Fusion Indicators												
Lastde	0.6533	0.6567	0.6967	0.4600	0.7367	**0.8300**	**0.9700**	0.8200	0.7200	**0.8467**	**0.7333**	**0.9667**
SpecFusion	0.6033	0.6233	0.7100	0.4300	0.7333	0.8167	0.9633	0.8367	0.6967	0.8033	**0.7333**	0.9333

Continuity and Stochasticity in Global Recasting When AI involvement expands into a continuous span, spectral evidence returns: in Complete, SpecDetect rises to 81.00% on GPT-4 and 97.00% on Llama-2 as autoregressive generation accumulates a low-variance machine pattern. Rewrite exposes the boundary, with Llama-2 Rewrite highly detectable by confidence metrics (LogRank 96.00%) while GPT-4 Rewrite is close to random for most metrics. Stable recasting is detectable, whereas high-entropy rewriting can reintroduce human-like roughness and weaken frequency-domain separation.

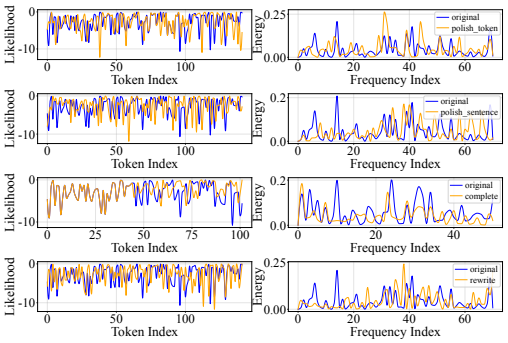


Figure 7: **Case Visualization of Signal Profiles across AI-Polishing Modes.** Original and polished signals under each mode. Left: time-domain log-likelihood; right: frequency-domain spectrum.

5.3.2 Humanizing: Injection and Destructive Disguise

Table 3 shows two strategies for altering machine signals in the AI $\rightarrow$ Human direction. The signal visualization is provided in Appendix D.8.

Mimicking Roughness via Conservative Adaptation Manual adaptation differentiates machine text by adding human-like fluctuations: SpecDetect rises from 67.67% on Adapt Token to 77.00% on Adapt Sentence. Broader intervention reintroduces human-authorship "grit", widening the gap from the smooth machine baseline.

Detectability of Destructive Disguise Humanize operations inject artificial perturbations such as typos or unnatural errors. These manipulations degrade text naturalness while disrupting training-free metrics with minimal effort, exposing detectors' reliance on statistical signatures over semantic coherence. This "less human" yet "less detectable" paradox exposes detector fragility.

Adaptive Confidence–Fluctuation Fusion Can adaptive weighting combine confidence and fluctuation more effectively across settings? We evaluate logistic scoring models that combine LogLikelihood and SpecDetect through interaction, length, and entropy features. Models are selected using labeled, grouped held-out clean documents, then frozen and applied to mixed-source and MixText settings. The entropy-adaptive model reaches 0.8409 on CoAuthor H-vs-L and 0.7632 on H-vs-C, compared with the respective strongest fixed baselines of 0.8138 and 0.7017. These gains do not extend to SemEval or MixText, indicating that weighting learned from clean documents transfers unevenly across operations. Appendix F describes the calibration protocol and transfer results.

5.4 Concluding Takeaways

The in-the-wild evaluation yields four points. **Scale Sensitivity**: short or fragmented units favor confidence indicators because spectral evidence has little context. **Locality vs. Continuity**: pointwise polishing is captured by local probability shifts, whereas continuous completion gives spectral indicators a coherent generated span. **The Deception Paradox**: humanizing and rewriting can distort confidence and fluctuation in different ways, so edit density is insufficient. **Metric Complementarity**: confidence and fluctuation provide complementary safety nets across mixed and edited text.

6 Conclusion

This work provides a unified explanation of what spectral detectors measure and systematically characterizes when confidence and fluctuation evidence succeeds or fails. Spectral scores reflect fluctuations in proxy probability trajectories, with their discriminative value depending on length, sampling scope, source continuity, and editing operation. Confidence cues complement this evidence in short, fragmented, and locally polished text. Adaptive weighting improves selected collaborative settings, while its transfer performance varies across operations. Together, these findings guide the use of confidence and fluctuation evidence according to the statistical signatures preserved in the text.

AI USE STATEMENT

In preparing this manuscript, we used generative AI tools only for language polishing. The authors developed the research ideas, designed and conducted the experiments, and take responsibility for the final manuscript.

REFERENCES

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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. *arXiv preprint arXiv:2310.05130*, 2023.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. Findings of the 2016 conference on machine translation (wmt16). In *First conference on machine translation*, pp. 131–198. Association for Computational Linguistics, 2016.

Ronald N Bracewell. The fourier transform. *Scientific American*, 260(6):86–95, 1989.

Evan N Crothers, Nathalie Japkowicz, and Herna L Viktor. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002, 2023.

Geoffrey M Currie. Academic integrity and artificial intelligence: is chatgpt hype, hero or heresy? In *Seminars in Nuclear medicine*, volume 53, pp. 719–730. Elsevier, 2023.

Holly Else. Abstracts written by ChatGPT fool scientists. *Nature*, 613:423, 2023.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*, 2019.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. Bias of ai-generated content: an examination of news produced by large language models. *Scientific Reports*, 14(1):5224, 2024.

David Faraggi and Benjamin Reiser. Estimation of the area under the roc curve. *Statistics in medicine*, 21(20):3093–3106, 2002.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*, 2019.

GH Hardy and EC Titchmarsh. A note on parseval's theorem for fourier transforms. *Journal of the London Mathematical Society*, 1(1):44–48, 1931.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*, 2019.

Ram Mohan Rao Kadiyala, Siddartha Pullakhandam, Kanwal Mehreen, Drishti Sharma, Siddhant Gupta, Jebish Purbey, Ashay Srivastava, Subhasya Tippareddy, Arvind Reddy Bobbili, Suraj Telugara Chandrashekhar, et al. Robust and fine-grained detection of ai generated texts. *arXiv preprint arXiv:2504.11952*, 2025.

Mina Lee, Percy Liang, and Qian Yang. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pp. 1–19, 2022.

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. *arXiv preprint arXiv:2508.11343*, 2025.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning*, pp. 24950–24962. PMLR, 2023.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*, 2018.

Andreas L Opdahl, Bjørnar Tessem, Duc-Tien Dang-Nguyen, Enrico Motta, Vinay Setty, Eivind Throndsen, Are Tverberg, and Christoph Trattner. Trustworthy journalism through ai. *Data & Knowledge Engineering*, 146:102182, 2023.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, 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. *arXiv preprint arXiv:2306.05540*, 2023.

Zhixiong Su, Yichen Wang, Herun Wan, Zhaohan Zhang, and Minnan Luo. Haco-det: A study towards fine-grained machine-generated text detection under human-ai coauthoring. *arXiv preprint arXiv:2506.02959*, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`, May 2021. Accessed: 2025-08-08.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, et al. Semeval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. *arXiv preprint arXiv:2404.14183*, 2024.

Clark Wissler. The spearman correlation formula. *Science*, 22(558):309–311, 1905.

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. *arXiv preprint arXiv:2410.06072*, 2024.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *arXiv preprint arXiv:2305.17359*, 2023.

Xuejun Yu, David J Nott, Minh-Ngoc Tran, and Nadja Klein. Assessment and adjustment of approximate inference algorithms using the law of total variance. *Journal of Computational and Graphical Statistics*, 30(4):977–990, 2021.

Zijie Zeng, Shiqi Liu, Lele Sha, Zhuang Li, Kaixun Yang, Sannyuya Liu, Dragan Gašević, and Guanliang Chen. Detecting ai-generated sentences in human-ai collaborative hybrid texts: challenges, strategies, and insights. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 7545–7553, 2024.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, et al. Llm-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 409–436, 2024.

APPENDIX

A EXTENDED RELATED WORK

A.1 TRAINING-FREE LLM TEXT DETECTION

Training-free detection identifies machine-generated text via intrinsic statistical signatures, categorized into distribution-based approaches Mitchell et al. (2023); Su et al. (2023); Yang et al. (2023); Bao et al. (2023) using perturbations and sample-based methods Solaiman et al. (2019); Su et al. (2023); Gehrmann et al. (2019); Ippolito et al. (2019); Xu et al. (2024); Luo et al. (2025) utilizing raw token statistics. We focus on the latter as they represent fundamental machine signal properties. Early research relied on confidence-based metrics like LogLikelihood, LogRank Solaiman et al. (2019), and Entropy Gehrmann et al. (2019); Ippolito et al. (2019), assuming LLMs exhibit higher statistical certainty than humans. Subsequent work like Lastde Xu et al. (2024) began modeling probability volatility, leading to the significant advancement of SpecDetect Luo et al. (2025). SpecDetect reframes detection as a signal processing task by analyzing fluctuations via the Discrete Fourier Transform (DFT). Despite its effectiveness, the statistical mechanism and operating conditions of frequency-domain detection remain insufficiently understood, a gap this study aims to bridge.

A.2 DETECTION IN COMPLEX SCENARIOS

Digital content increasingly involves complex scenarios like mixed-source text Lee et al. (2022); Wang et al. (2024) and collaborative authorship Zhang et al. (2024), where interleaved segments and human-AI editing alter the text's statistical traces. While recent studies Zeng et al. (2024); Su et al. (2025); Kadiyala et al. (2025) employ supervised models for sequence labeling or boundary detection, they often suffer from limited cross-domain generalization Zeng et al. (2024). Conversely, training-free detection remains confined to long-form, single-source documents, leaving its performance in intricate scenarios unexplored. Although frequency-domain methods such as SpecDetect Luo et al. (2025) effectively capture generative fluctuations at the document level, it remains unknown whether indicators can maintain discriminative power when signals are fragmented by mixing or masked by collaborative editing. This work interrogates these complex regimes to define the utility and limitations of frequency-domain detection beyond the traditional long-form, single-source paradigm.

B MATHEMATICAL PROOFS

B.1 DERIVATION OF THEOREM 1

Proof. For a text source s and token position t, the indicator $b_t^{(s)}$ records whether the token falls in $\mathcal{V}_{\text{tail}}^{(t)}(\tau)$, and γ_s is the average rate at which this event occurs. Applying the law of total variance Yu et al. (2021) to the mixture signal $x_s[t]$ gives

$$\text{Var}(x_s[t]) = \mathbb{E}[\text{Var}(x_s[t] \mid b_t^{(s)})] + \text{Var}(\mathbb{E}[x_s[t] \mid b_t^{(s)}]). \tag{5}$$

The first term combines the conditional variances inside the head and tail regions:

$$\mathbb{E}[\text{Var}(x_s[t] \mid b_t^{(s)})] = (1-\gamma_s)\sigma_{\text{head}}^2 + \gamma_s\sigma_{\text{tail}}^2. \tag{6}$$

The second term captures the variance caused by switching between the head and tail means:

$$\text{Var}(\mathbb{E}[x_s[t] \mid b_t^{(s)}]) = \gamma_s(1-\gamma_s)(\mu_{\text{tail}} - \mu_{\text{head}})^2. \tag{7}$$

Combining the two terms gives the group-wise variance

$$\begin{aligned}\sigma_s^2 = (1-\gamma_s)\sigma_{\text{head}}^2 + \gamma_s\sigma_{\text{tail}}^2 \\ + \gamma_s(1-\gamma_s)(\mu_{\text{tail}} - \mu_{\text{head}})^2.\end{aligned} \tag{8}$$

Substituting $s = H$ and $s = \text{AI}$ and subtracting σ_{AI}^2 from σ_H^2 yields

$$\begin{aligned}\Delta\sigma^2 = (\gamma_H - \gamma_{\text{AI}})(\sigma_{\text{tail}}^2 - \sigma_{\text{head}}^2) \\ + \big[\gamma_H(1-\gamma_H) \\ - \gamma_{\text{AI}}(1-\gamma_{\text{AI}})\big](\mu_{\text{tail}} - \mu_{\text{head}})^2.\end{aligned} \tag{9}$$

Additionally, in the idealized white-box truncation case where generated tokens remain in the head region, setting $\gamma_{\mathrm{AI}} = 0$ gives

$$\Delta\sigma^2 = \gamma_H(\sigma^2_{\mathrm{tail}} - \sigma^2_{\mathrm{head}}) + \gamma_H(1-\gamma_H)(\mu_{\mathrm{tail}} - \mu_{\mathrm{head}})^2. \tag{10}$$

The human–machine fluctuation gap therefore emerges when the tail rate and conditional log-probability structure differ across text sources. Tail-region log-probabilities are typically lower and more dispersed, so a higher human tail-token rate increases the variance gap under the stated shared-statistics view. □

B.2 Proof of Corollary 2

Proof. For a text source $s \in \{H, \mathrm{AI}\}$, let $z_s[t] = x_s[t] - \mu_s$ be the centered scoring-model log-probability sequence of length N. We use the Discrete Fourier Transform (DFT) Bracewell (1989),

$$\hat{z}_k^{(s)} = \sum_{t=0}^{N-1} z_s[t] e^{-j\frac{2\pi}{N}kt}. \tag{11}$$

Parseval's theorem Hardy & Titchmarsh (1931) for this convention states:

$$\sum_{k=0}^{N-1} |\hat{z}_k^{(s)}|^2 = N \sum_{t=0}^{N-1} z_s[t]^2. \tag{12}$$

We define the average spectral energy as

$$\bar{\mathcal{E}}_s = \frac{1}{N} \sum_{k=0}^{N-1} |\hat{z}_k^{(s)}|^2. \tag{13}$$

Substituting Parseval's identity into this definition gives

$$\bar{\mathcal{E}}_s = \sum_{t=0}^{N-1} z_s[t]^2. \tag{14}$$

The variance of the centered signal is

$$\sigma_s^2 = \frac{1}{N} \sum_{t=0}^{N-1} z_s[t]^2, \quad \sum_{t=0}^{N-1} z_s[t]^2 = N\sigma_s^2. \tag{15}$$

Therefore,

$$\bar{\mathcal{E}}_s = N\sigma_s^2. \tag{16}$$

Taking expectations for human and AI text sources and subtracting gives

$$\Delta\bar{\mathcal{E}} = \mathbb{E}[\bar{\mathcal{E}}_H] - \mathbb{E}[\bar{\mathcal{E}}_{\mathrm{AI}}] = N(\sigma_H^2 - \sigma_{\mathrm{AI}}^2) = N\Delta\sigma^2. \tag{17}$$

Thus, the spectral-energy gap follows directly from the time-domain fluctuation divergence established in Theorem 1. □

C Experimental Setup Details

C.1 Metric Details

In this section, we provide the formal definitions for all detection metrics. For consistency, all metrics are formulated as a score function $\mathcal{S}$ where a **higher value indicates a higher probability of machine generation.**

C.1.1 CONFIDENCE-BASED METRICS

These metrics measure the average probability intensity and capture the mean shift inherent in AI's restricted sampling.

- **LogLikelihood (LL)** Solaiman et al. (2019): The primary measure of model certainty. Since AI models maximize likelihood, machine text yields higher values.

$$\mathcal{S}_{\text{LL}} = \frac{1}{N}\sum_{t=1}^{N} \log P(w_t|w_{<t}) \tag{18}$$

- **LogRank (LR)** Solaiman et al. (2019): The rank-domain counterpart to Log-Likelihood. It measures the usage of high-probability (low-rank) tokens. To align polarity, we use the negative logarithmic rank:

$$\mathcal{S}_{\text{LR}} = -\frac{1}{N}\sum_{t=1}^{N} \log(\text{Rank}(w_t)) \tag{19}$$

- **LRR (Log-Likelihood Ratio)** Su et al. (2023): A normalized metric that balances probability magnitude against rank magnitude. It is defined directly as the ratio of Log-Likelihood to the average Log-Rank:

$$\mathcal{S}_{\text{LRR}} = \frac{\text{LogLikelihood}}{\text{LogRank}} = \frac{\frac{1}{N}\sum_{t=1}^{N} \log P(w_t)}{\frac{1}{N}\sum_{t=1}^{N} \log(\text{Rank}(w_t))} \tag{20}$$

- **Entropy** Gehrmann et al. (2019); Ippolito et al. (2019): Captures predictive uncertainty. AI text exhibits low entropy (high certainty). We define the score as the negative average entropy:

$$\mathcal{S}_{\text{Entropy}} = -\frac{1}{N}\sum_{t=1}^{N} \mathbb{H}(P(\cdot|w_t)) \tag{21}$$

C.1.2 FLUCTUATION-BASED METRICS

These metrics measure the structural variance or spectral density of the signal.

- **SpecDetect** Luo et al. (2025): Integrates the spectral energy of the log-probability signal. Smaller log-probability fluctuations yield lower spectral energy. To account for differences in sequence length across candidate samples, we adopt the average spectral energy.
  Let the original time-domain signal be $x[t] = \log P(w_t|w_{<t})$. We first perform zero-centering to remove the mean component:

$$\hat{x}[t] = x[t] - \frac{1}{N}\sum_{j=1}^{N} x[j]. \tag{22}$$

  We then define the score as the negative average power of the DFT coefficients $X[k]$:

$$\mathcal{S}_{\text{SpecDetect}} = -\frac{1}{N}\sum_{k=0}^{N-1} |X[k]|^2, \tag{23}$$

  where $X[k] = \sum_{t=0}^{N-1} \hat{x}[t] \cdot e^{-i\frac{2\pi}{N}kt}$.

C.1.3 FUSION-BASED METRICS

These indicators combine the mean (confidence) and variance (fluctuation) dimensions to bridge the blind spots of single-view metrics.

- **Lastde** Xu et al. (2024): This method models fluctuation using Multiscale Diversity Entropy (MDE). The score is defined as the ratio of likelihood to fluctuation. It reaches its maximum when a signal is simultaneously high-confidence and low-fluctuation:

$$\mathcal{S}_{\text{LastDe}} = \frac{\mathcal{S}_{\text{LL}}}{\text{MDE}} \tag{24}$$

Table 4: **Dataset Statistics.** $Avg.\ Len.$ is the word count per detection unit. $Mixed$ denotes mixed-source text within a document; $Collab$ indicates collaborative content.

Dataset	Sub-task	LLM Engine	Samples	Avg. Len.	Level	Mixed	Collab
SemEval	Prefix-Comp.	GPT-3	4,154	19.76	Sent.	✓	×
CoAuthor	Interleaved	GPT-3.5	1,445	14.71	Sent.	✓	✓
MixText	AI-Polishing	GPT-4, Llama-2	3,000	130.18	Doc.	×	✓
	Humanizing	GPT-4, Llama-2	1,500	128.66	Doc.	×	✓

where MDE denotes the magnitude of Multiscale Diversity Entropy computed on the rank sequence. Adopting the default configuration, we set the hyperparameters as follows: $s = 3$, $\epsilon = 10 \times N$, and $\tau^{'} = 5$, with N representing the number of tokens in the text.

- **SpecFusion**: A standardized linear combination. Since both $\mathcal{S}_{\text{LL}}$ and $\mathcal{S}_{\text{SpecDetect}}$ are aligned such that higher values indicate AI, we sum their standardized scores:

$$\mathcal{S}_{\text{Fusion}} = Z(\mathcal{S}_{\text{LL}}) + Z(\mathcal{S}_{\text{SpecDetect}}) \tag{25}$$

  where $Z(\cdot)$ denotes Z-score standardization. The normalization scope $\mathcal{C}$ depends on the task granularity:
  - For document-level detection, $\mathcal{C}$ includes all documents in the evaluation corpus.
  - For sentence-level detection, $\mathcal{C}$ includes all sentences within the current document, effectively normalizing against the local context.

C.2 Dataset Details

To evaluate detection performance across different granularities of human-AI interaction, the experiments are categorized into two primary real-world scenarios. Table 4 summarizes the statistics of the datasets utilized in this study. The selected benchmarks cover diverse detection levels (i.e., sentence-level and document-level) and originate from various LLM generators (e.g., GPT-3/4 Achiam et al. (2023), Llama-2 Touvron et al. (2023)), ensuring a comprehensive assessment of metric robustness under varying signal densities.

Scenario I: Mixed-Source Text. This scenario targets documents where human-authored and AI-generated sentences coexist, reflecting common "spliced" or "interleaved" content structures found in news co-writing or academic stitching.

- **SemEval-2024 (Task 8)** Wang et al. (2024): The experiments utilize Subtask C from this benchmark. The data consists of documents with a human-written prefix followed by a machine-generated continuation (GPT-3). Since the transition point is variable, this task evaluates the capability to identify machine segments within a composite context.
- **CoAuthor** Lee et al. (2022): Moving beyond the simple prefix-continuation structure, CoAuthor simulates a complex environment where sentences are interleaved throughout the text. Each document contains sentences from three sources: **Human**, **LLM**, and **Collaborative**. This setup serves to evaluate detection performance in non-sequential mixing environments.

Scenario II: Collaborative Text. Complementing the structural mixing in CoAuthor, the **MixText** Zhang et al. (2024) dataset is employed to conduct a granular investigation into specific editing operations. The dataset comprises 300 sample pairs for each modification direction, generated using GPT-4 and Llama-2-70b engines.

- **AI-Polishing (Human → AI)**: Machine-led refinement of human drafts. This category includes three specific modes:
  - *Polish*: Involves **Token-level** and **Sentence-level** refinement. In this mode, the LLM focuses on enhancing fluency and correcting minor errors while strictly adhering to the original semantic skeleton.

  - *Rewrite*: The LLM extracts key information and regenerates the text entirely. Unlike simple polishing, this operation significantly alters the syntactic structure.
  - *Complete*: The LLM generates the remaining 2/3 of the text based on a human-authored 1/3 prefix, representing a continuous autoregressive generation process.
- **Humanizing (AI → Human)**: Strategies used to disguise machine origin, categorized into two opposing approaches:
  - *Humanize*: A "destructive" strategy that introduces artificial noise (e.g., typos, grammatical errors) to mimic human writing imperfections. This injection process is specifically implemented using prompt-engineering on GPT-4 and Llama-2 models.
  - *Adapt*: A "conservative" strategy where humans align machine text with natural linguistic habits to enhance fluency. Similar to the polishing task, this operation is further stratified into **Token-level** and **Sentence-level** adaptations to verify detection sensitivity to edit granularity.

C.3 Implementation Details

Following the configurations established in our document-level experiments, we utilize **GPT-J-6B** as the base proxy model to extract all statistical metrics. To address the unique challenges posed by diverse real-world human-AI interactions, we adopt two distinct evaluation protocols:

- **Scenario I: Mixed-Source Text Evaluation**. This protocol measures the ability to distinguish between authorship sources within a single, heterogeneous document context. We perform sentence-level scoring across the document and report the average AUC calculated within each individual sample. This approach ensures that the detection performance reflects the metric's sensitivity to fragmented signals and local statistical shifts within each document.
- **Scenario II: Collaborative Text Evaluation**. Focusing on the MixText dataset, we employ pair-wise accuracy to assess detection performance. This metric calculates the percentage of samples where the metric score correctly shifts following a specific editing operation. For instance, a successful detection is recorded when an AI-polished version of a text receives a higher detection score than its original human-authored counterpart. This methodology directly measures the sensitivity of each indicator to the fine-grained statistical transformations introduced by collaborative editing and humanization processes.

All experiments are conducted on a high-performance computing node equipped with a single 80GB NVIDIA H800 GPU. To ensure the reproducibility of our results, our implementation builds upon the open-source frameworks provided by the original authors. The complete codebase and configurations are available at the following repositories:

- **Other metrics**: `https://github.com/TrustMedia-zju/Lastde_Detector`
- **SpecDetect**: `https://github.com/luohaitong/SpecDetect`

D Additional Experiment Results

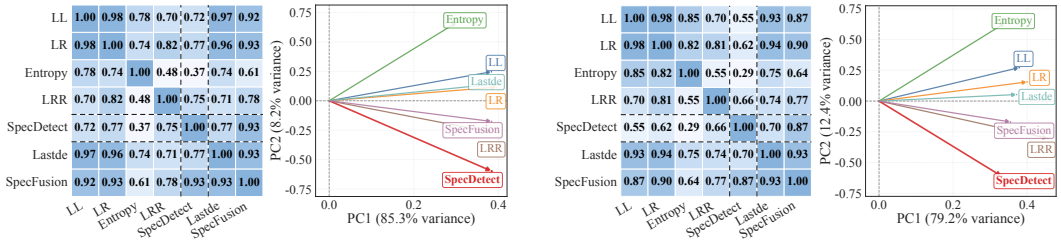
D.1 Metric Distinctness Validation

To validate the generality of our findings, we extend the dimensionality analysis to the XSum and Reddit datasets.

As shown in Figure 8, the results are consistent with the Writing dataset results presented in the main text. The correlation heatmaps confirm that metrics cluster tightly within their respective families. Similarly, the PCA loading plots demonstrate that spectral methods capture a distinct signal dimension separate from the confidence-based probability.

D.2 Input Signal Granularity Investigation

To capture spectral signatures accurately, the choice of the input proxy signal $x[t]$ is critical. While prior work Mitchell et al. (2023); Bao et al. (2023) demonstrates that LogRank provides a robust, scale-invariant measure for Confidence-based metrics, whether this preference holds for fluctuation-based indicators remains unexamined.



(a) Results on XSum Dataset (b) Results on Reddit Dataset

Figure 8: **Analysis of Metric Relation on Additional Datasets.** (Left) **Correlation Heatmap**: Metrics cluster into Confidence and Fluctuation families based on Spearman correlation. (Right) **PCA Loading Plot**: Illustrates the geometric relationship where PC1 captures the dominant variance, while PC2 captures the secondary direction of variation.

As illustrated in Figure 9, we observe a distinct divergence in signal preference across three datasets. While confidence-based metrics improve by $+2\% \sim 4\%$ AUC with LogRank, fluctuation-based metrics suffer a performance collapse, achieving optimal results *only* with continuous LogLikelihood.

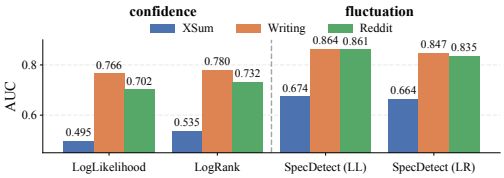


Figure 9: **Input Proxy Ablation across All Metrics.** While confidence metrics benefit from discretized LogRank, fluctuation-based metrics require continuous LogLikelihood to resolve high-frequency spectral signatures. Comparative analysis of input signal preference. While confidence metrics improve with discretized LogRank inputs, fluctuation-based metrics (i.e., SpecDetect) suffer a severe performance collapse, confirming the necessity of continuous LogLikelihood for resolving micro-fluctuations.

The difference reflects the information retained by each signal. LogRank records a token's position in the probability ordering, whereas LogLikelihood retains its probability magnitude. Ranking can therefore change the fluctuations observed along the trajectory. In these experiments, LogLikelihood yields stronger spectral detection, while mean LogRank remains competitive as a confidence-based baseline.

D.3 Complete Main-Setting Sensitivity Results

This subsection complements the main setting, where Llama2-13B is the source model and GPT-J-6B is the proxy scorer. The main text highlights representative metrics and datasets; here we report fuller curves for length and sampling sensitivity.

D.3.1 Length Sensitivity

Figure 10 reports the all-metric length curves under the main source/proxy setting. The result is consistent with Figure 5: fluctuation-based metrics gain more clearly as sequence length increases, because longer spans provide more evidence for stable variance patterns.

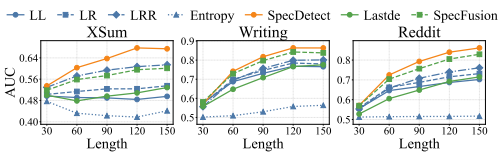
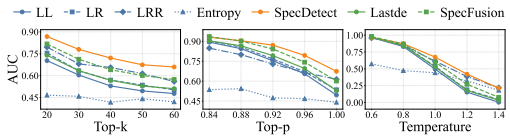


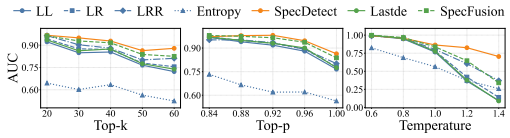
Figure 10: **Performance of All Metrics vs. Sequence Length (Llama-2 Source).** Fluctuation-based metrics achieve more significant gains as sequence length increases.

D.3.2 Sampling Scope

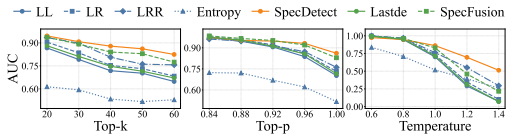
Figure 11 reports the all-dataset sampling-scope curves under the main source/proxy setting. The pattern follows Figure 6: broader sampling generally weakens separability, while tighter decoding makes the machine signal easier to distinguish.



(a) Results on XSum Dataset



(b) Results on writing Dataset



(c) Results on Reddit Dataset

Figure 11: **Impact of Decoding Strategies across All Datasets (Llama-2 Source).** Tighter sampling generally improves detection across the three datasets, consistent with reduced fluctuation in machine-generated text.

D.4 Additional Source-Model Sensitivity Results

This subsection checks whether the length and sampling-scope observations persist when the source model changes. We use GPT-4-Turbo for the additional length study and Qwen-3-8B for the additional sampling-scope study. The results are used as qualitative source-model checks, since the exact AUC values and slopes naturally depend on the generator.

D.4.1 Length Sensitivity

Figure 12 shows the length-sensitivity curves for GPT-4-Turbo generated text. The trend remains aligned with the main setting: SpecDetect initially trails on some datasets but grows strongly with longer observation windows.

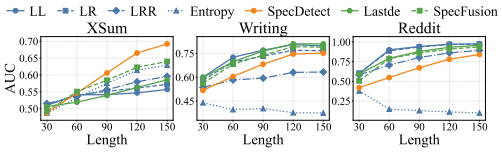
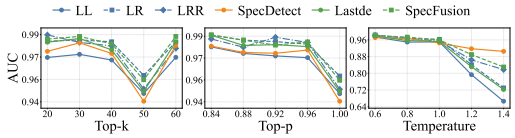


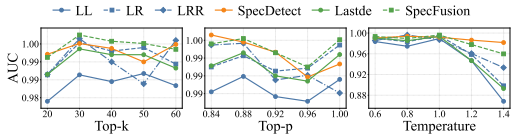
Figure 12: **Performance of All Metrics vs. Sequence Length (GPT-4 Source).** While SpecDetect initially trails on Writing and Reddit, it achieves the steepest growth among all metrics as length increases, confirming that spectral methods require sufficient observation windows to resolve stable variance.

D.4.2 Sampling Scope

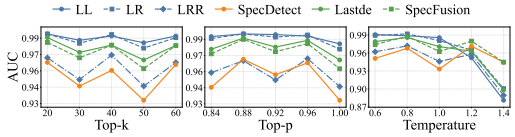
Figure 13 shows the sampling-scope curves for Qwen-3-8B generated text. The performance gradient is more subtle under this source model, but the overall trajectories remain consistent with the main conclusion that stricter sampling improves separability. We exclude Entropy from these visualizations for clarity because its near-random performance otherwise obscures the more discriminative indicators.



(a) Results on XSum Dataset



(b) Results on writing Dataset



(c) Results on Reddit Dataset

Figure 13: **Impact of Decoding Strategies across All Datasets (Qwen-3-8B Source).** Tighter sampling improves separability across the three evaluation scenarios, while the exact gradients vary with the source model.

D.5 Additional Head/Tail Validation

These additional head/tail checks are consistent with the main-text GPT-J-6B Tail@0.90 result in Figure 3. Figure 14 uses GPT-J-6B with a Rank>50 threshold, while Figures 15 and 16 repeat the Tail@0.90 and Rank>50 checks using Llama3-8B. In all three views, human continuations enter the proxy tail region more often than paired AI continuations, with larger gaps on Reddit and WritingPrompts and a smaller gap on XSum.

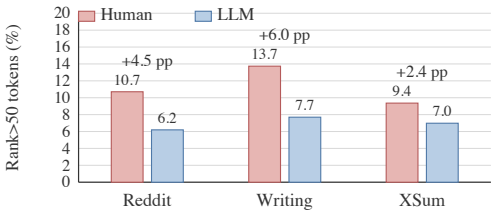


Figure 14: **GPT-J-6B Rank>50 head/tail check.** Bars show the percentage of tokens whose proxy-model rank is greater than 50.

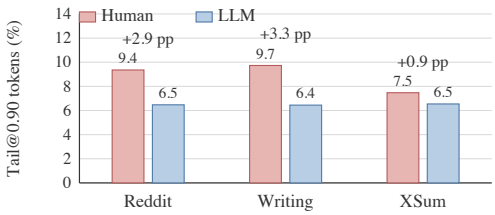


Figure 15: **Llama3-8B Tail@0.90 head/tail check.** Bars show the percentage of tokens outside the Top-$p = 0.90$ head set under Llama3-8B proxy scoring.

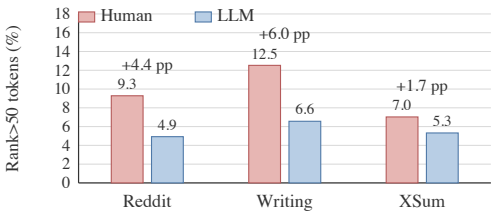


Figure 16: **Llama3-8B Rank>50 head/tail check.** Bars show the percentage of tokens whose proxy-model rank is greater than 50.

D.5.1 HEAD/TAIL CONTRIBUTIONS TO LOG-PROBABILITY VARIANCE

For each source $s \in \{H, \mathrm{AI}\}$, log-probability variance combines variation within the head and tail regions with the separation between their means. Let $\mu_{r,s}$ and $\sigma^2_{r,s}$ denote the conditional mean and variance in region $r \in \{\mathrm{head}, \mathrm{tail}\}$ for source s. The law of total variance gives

$$\sigma_s^2 = (1-\gamma_s)\sigma^2_{\mathrm{head},s} + \gamma_s\sigma^2_{\mathrm{tail},s} + \gamma_s(1-\gamma_s)(\mu_{\mathrm{tail},s} - \mu_{\mathrm{head},s})^2. \tag{26}$$

With shared conditional moments, subtracting the two source variances recovers Theorem 1.

We compute this decomposition separately for each text using GPT-J-6B and the Top-$p = 0.90$ partition, with 150 matched human–LLM pairs per dataset. The mean paired human-minus-LLM log-probability variance gaps are 0.8168 on XSum, 2.3081 on WritingPrompts, and 1.8317 on Reddit; all paired-bootstrap 95% confidence intervals lie above zero under the protocol in Appendix E. Thus, the higher human fluctuation observed in these datasets persists when the conditional moments are estimated separately for each source.

D.6 LLAMA3-8B PROXY RESULTS FOR MAIN EXPERIMENTS

This section reports the available Llama3-8B scoring-proxy counterparts for the main experiments. These figures and tables complement the GPT-J-6B proxy results in the main text. Overall, the qualitative conclusions are consistent with the GPT-J-6B setting: longer spans strengthen fluctuation-based evidence, broader sampling weakens all detectors while preserving a relative fluctuation advantage in the evaluated range, and mixed or collaborative sentence-level settings remain harder than pure H-vs-L detection.

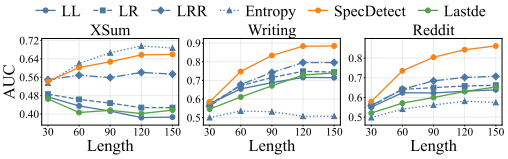


Figure 17: **Llama3-8B proxy length sensitivity.** This figure mirrors Figure 5 with Llama3-8B as the scoring proxy. The qualitative pattern is consistent with the main text: longer spans tend to strengthen fluctuation-based evidence, while exact values vary with the proxy model.

D.6.1 LENGTH SENSITIVITY

Figure 17 shows the Llama3-8B proxy counterpart of the main length-sensitivity analysis. The qualitative trend is consistent with Figure 5: longer spans usually strengthen fluctuation-based evidence, although the exact AUC values and slopes vary with the proxy model.

D.6.2 SAMPLING SCOPE

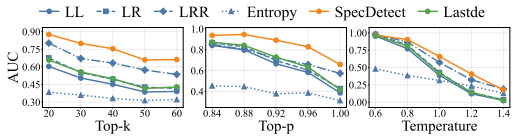
Figure 18 shows the Llama3-8B proxy counterpart of the sampling-scope analysis across XSum, WritingPrompts, and Reddit. The pattern is aligned with Figure 6: broader sampling generally lowers detector performance, while fluctuation-based indicators show relatively smoother degradation in the evaluated range.

D.6.3 SENTENCE-LEVEL PURE AND COLLABORATIVE TEXT

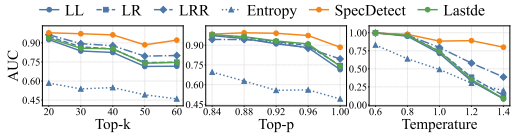
Table 5 reports the Llama3-8B proxy counterpart of Table 2. The results follow the same broad pattern: confidence metrics are strong in pure H-vs-L sentence-level settings, while collaborative rows are harder and fluctuation evidence becomes more useful in the L-vs-C comparison.

D.6.4 MIXTEXT PAIRWISE ACCURACY

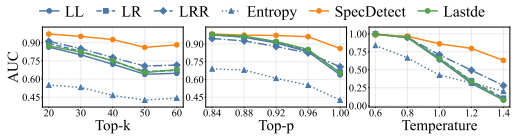
Table 6 reports the Llama3-8B proxy counterpart of Table 3. The operation-level pattern is similar to the main text: continuous completion favors fluctuation or hybrid indicators, while local polishing and humanizing operations often rely more on confidence shifts.



(a) Results on XSum Dataset



(b) Results on WritingPrompts Dataset



(c) Results on Reddit Dataset

Figure 18: **Impact of Decoding Strategies across All Datasets (Llama3-8B Proxy).** This figure mirrors the sampling-scope analysis with Llama3-8B as the scoring proxy. Broader sampling generally weakens separability, while exact curves vary with the proxy model and dataset.

Table 5: **Llama3-8B Proxy Sentence-level Detection across Pure and Collaborative Text.** Average within-document AUC for binary classification among Human (H), LLM (L), and Collaborative (C) sentences, following the same format as Table 2. Best results are **bolded**; second-best are underlined.

Metric	Pure Generation		Collaborative Mixed	
	SemEval	CoAuthor (H vs. L)	CoAuthor (H vs. C)	CoAuthor (L vs. C)
Confidence-based				
LogLikelihood	**0.8698**	0.6415	0.7083	0.5839
LogRank	0.8449	**0.6483**	**0.7221**	0.5982
LRR	0.2190	0.5152	0.5152	0.5276
Entropy	0.2981	0.5306	0.5306	0.4751
Fluctuation-based				
SpecDetect	0.7141	0.5639	0.6721	**0.7479**
Fusion Indicator				
SpecFusion	0.8479	0.6412	0.6931	0.7453

D.7 Cross-Language Test

This appendix reports a WMT16 Bojar et al. (2016)(German) cross-language test for the two boundary conclusions in the main text. We provide a length-sensitivity view and a decoding-scope view following the main protocol.

Figure 19 and 20 summarizes the WMT16 check. The two figures report length sensitivity and decoding-scope sensitivity for the same dataset.

Table 6: **Llama3-8B Proxy Pairwise Accuracy on MixText Scenarios.** Values are the percentage of pairs where the Modified version is ranked more machine-like than the Original, following the same column order and formatting as Table 3. Best results are **bolded**; second-best are underlined.

Metric	AI-Polishing (GPT-4)				AI-Polishing (Llama-2)				Humanizing			
	Polish Token	Polish Sentence	Complete (1/3+2/3)	Rewrite (Re-gen)	Polish Token	Polish Sentence	Complete (1/3+2/3)	Rewrite (Re-gen)	Adapt Token	Adapt Sentence	Llama-2 Humanize	GPT-4 Humanize
Confidence-based Metrics												
LogLikelihood	0.3733	**0.6267**	0.5667	0.5300	0.6933	0.8133	0.9633	0.7733	**0.7600**	**0.8833**	**0.7267**	**0.9900**
LogRank	0.5033	0.5167	0.6467	0.3767	0.6833	0.7867	0.9767	0.7567	0.7200	0.8600	0.7100	0.9733
LRR	0.3200	0.3367	0.5100	0.5333	0.3133	0.2100	0.1067	0.2600	0.1467	0.1267	0.2967	0.0433
Entropy	0.3533	0.4300	0.5567	**0.6233**	0.2567	0.2100	0.1133	0.2100	0.1667	0.1600	0.2433	0.0300
Fluctuation-based Metrics												
SpecDetect	**0.5633**	0.6133	**0.8033**	0.4900	0.6933	0.8100	0.9600	0.7633	0.6967	0.8433	0.6600	0.9067
Hybrid Fusion Indicators												
Lastde	0.5500	0.5933	0.6633	0.3933	**0.7233**	0.8200	**0.9800**	**0.8033**	0.7433	0.8500	0.6933	0.9833
SpecFusion	0.5533	0.5833	0.7433	0.3967	0.7100	**0.8267**	**0.9800**	0.7733	0.7200	0.8600	0.6967	0.9600

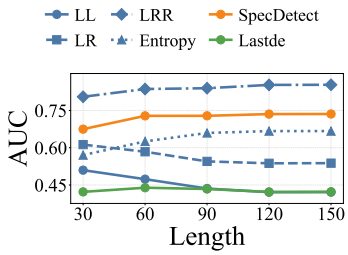


Figure 19: **WMT16 cross-language test (length sensitivity).** Following Figure 5.

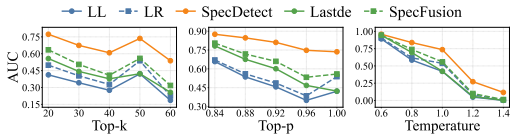


Figure 20: **WMT16 cross-language test (decoding-scope sensitivity).** Following Figure 6.

D.8 MixText Edit-Density Diagnostics

Edit density measures how much a modified text changes relative to its paired original. We tokenize both versions and count token-level insertions, deletions, and substitutions, with replacements counted by the larger replaced span:

$$\text{EditDensity}(x, \tilde{x}) = \frac{\text{EditOps}(x, \tilde{x})}{\max(|x|, |\tilde{x}|)}.$$

A value near 0 indicates light local editing, while values near or above 1 indicate near-full rewriting, destructive humanizing, or substantial length expansion.

To test whether this quantity predicts detector behavior, we compare per-sample edit density with signed detector-score movement. Let $S_m(\cdot)$ denote the score of metric m, where larger values indicate stronger machine evidence. For pair j in task d, we define

$$\Delta^{(d)}_{j,m} = \eta_d\big(S_m(\tilde{x}_j) - S_m(x_j)\big),$$

where $\eta_d = +1$ for Human→AI operations and $\eta_d = -1$ for AI→Human operations, so positive values follow the expected pairwise direction in Table 3. For each task–metric pair, we compute the Spearman correlation $\rho_{d,m}$ between edit density and $\Delta^{(d)}_{j,m}$ over the 300 paired samples, and then average $\rho_{d,m}$ across the 12 MixText tasks for each metric. The largest metric-level mean correlation is only 0.127, indicating that edit density alone weakly explains detector-score movement.

Table 7 reports the average edit density for all 12 MixText tasks, together with the best-performing main-text metric as a compact reference to Table 3.

Setting	Direction	Operation	Mean density	Best metric	Pair acc.
GPT-4	H→AI	Polish-token	0.336	Entropy	0.7867
GPT-4	H→AI	Polish-sentence	0.800	Entropy	0.7067
GPT-4	H→AI	Complete	1.089	SpecDetect	0.8100
GPT-4	H→AI	Rewrite	1.051	LRR	0.4933
Llama-2	H→AI	Polish-token	0.635	Entropy	0.8233
Llama-2	H→AI	Polish-sentence	0.810	LogRank / Lastde	0.8300
Llama-2	H→AI	Complete	0.958	SpecDetect / Lastde	0.9700
Llama-2	H→AI	Rewrite	0.923	LogRank	0.9600
Human	AI→H	Adapt-token	0.098	Entropy	0.7733
Human	AI→H	Adapt-sentence	0.228	Lastde	0.8467
Llama-2	AI→H	Humanize	0.985	LogLikelihood / LogRank / Lastde / SpecFusion	0.7333
GPT-4	AI→H	Humanize	0.981	LogLikelihood / Lastde	0.9667

Table 7: **Edit density and compact detection summary for all MixText operations.** Mean density is the average token-level edit density over 300 paired samples. The best metric and pairwise accuracy are from Table 3.

The table shows why edit density is informative but insufficient as a collapse threshold. Complete operations introduce sustained machine generation and favor fluctuation-based evidence, while GPT-4 rewrite has similarly high density but remains difficult. Low-density adaptation and polishing can still be detected by confidence shifts, and humanize operations reflect injected noise and edit direction rather than density alone.
Figure 21 visualizes representative time- and frequency-domain changes for the Humanizing operations discussed in Section 5.3.2.

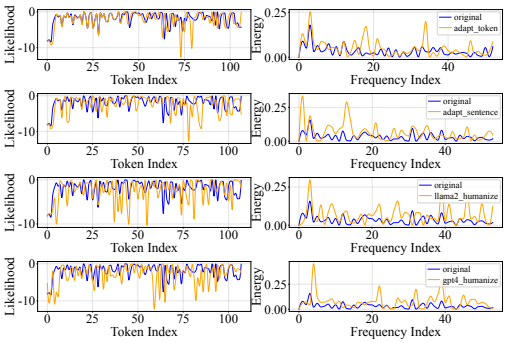


Figure 21: **Case Visualization of Signal Profiles across Humanizing Modes.** Original and humanized signals under each mode. Left: time-domain log-likelihood; right: frequency-domain spectrum.

E Statistical Reliability of the Empirical Findings

Protocol. We assess sampling uncertainty using 2,000 bootstrap resamples of fixed detector scores. Resampling preserves matched human–LLM pairs for document-level AUC, documents/sessions for mixed-source mean within-document AUC, and aligned original–modified pairs for MixText pairwise accuracy. Method comparisons use the same sampled units to compute pointwise percentile 95% confidence intervals (CIs) for the score differences. The analysis covers the reference setting, response lengths, Top-p, Top-k, temperature, mixed-source comparisons, and MixText operations.

Findings. SpecDetect shows greater length-related gains than LogLikelihood across all three document-level datasets: paired-bootstrap intervals for the differences in fitted AUC slopes over 30–150 words lie above zero. Higher-temperature sampling reduces spectral discriminability: on each dataset, SpecDetect has lower AUC at $T = 1.4$ than at $T = 0.8$, with non-overlapping 95% intervals. In mixed-source detection, LogLikelihood and LogRank outperform SpecDetect on SemEval and CoAuthor H-vs-L, while the ordering reverses on CoAuthor H-vs-C.

Table 8: Representative paired differences (SpecDetect minus LogLikelihood). Positive values favor SpecDetect; negative values favor LogLikelihood.

Setting	Difference
Document-level (AUC)	
XSum	$+0.1789$
Writing	$+0.0981$
Reddit	$+0.1592$
MixText (pairwise accuracy)	
GPT-4 token polish	-0.1233
GPT-4 complete	$+0.1600$

Representative comparisons. Table 8 contrasts the reference 150-word setting with two editing operations. SpecDetect outperforms LogLikelihood on the three document-level datasets. GPT-4 token polishing instead favors LogLikelihood, while continuous completion favors SpecDetect. The paired-bootstrap 95% intervals for all five comparisons exclude zero. Full bootstrap outputs accompany the result files.

F Adaptive Confidence–Fluctuation Fusion

Features and calibration. We combine standardized LogLikelihood z_C and SpecDetect z_F with log token length z_ℓ and negative mean predictive entropy z_E. Logistic scores take the form

$$g(x) = b + w_C z_C + w_F z_F + w_{CF} z_C z_F + \sum_{u \in \{\ell, E\}} (w_{Cu} z_C z_u + w_{Fu} z_F z_u).$$

We compare seven feature families: linear, interaction, length, length–interaction, entropy, length–entropy, and length–entropy–interaction, each with L2 inverse regularization in $\{0.01, 0.1, 1, 10\}$. Five-fold grouped cross-validation on clean documents from the three datasets selects models by mean condition-level AUC, breaking ties by worst-condition AUC. Each matched pair and its generation conditions stay in the same fold, with groups defined by dataset and original pair ID. Standardization and fitting use only training folds.

Within the four feature families without entropy, cross-validation selects the interaction model. Across all seven families, it selects the length–entropy–interaction model, termed entropy-adaptive here. Both use inverse regularization 10, are refit on all labeled clean calibration data, and are applied unchanged across target operations. We compare these calibrated models with the fixed training-free baselines. Candidate scores, fitted parameters, and standardization constants are provided in the accompanying reproducibility files.

Transfer behavior. In Figure 22, entropy-adaptive fusion improves CoAuthor H-vs-L from 0.8138 to 0.8409 and H-vs-C from 0.7017 to 0.7632, but falls below the strongest fixed baselines on SemEval, CoAuthor L-vs-C, and MixText. Interaction fusion stays closer to the fixed baselines without consistent gains. The benefit of clean-data calibration therefore varies across target operations.

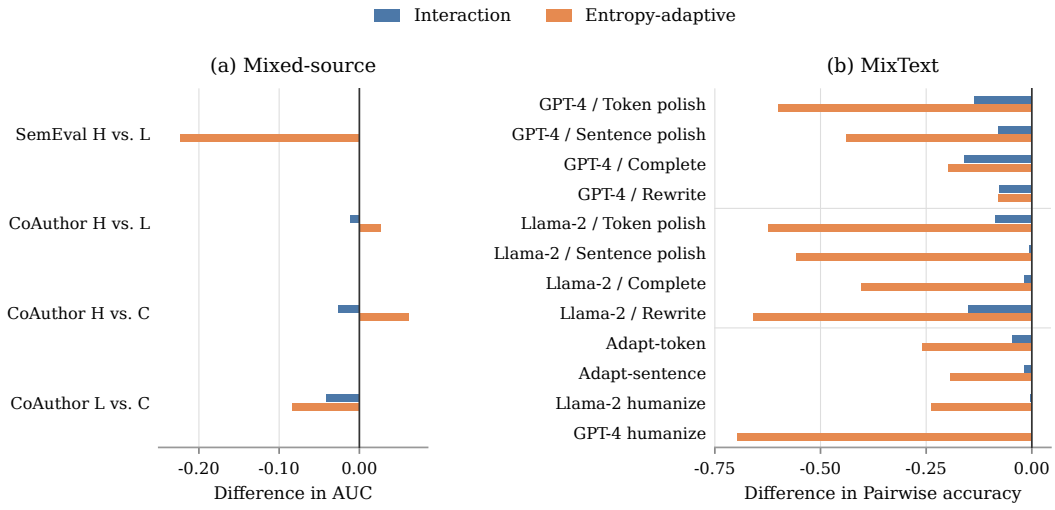


Figure 22: **Cross-operation transfer of adaptive fusion.** Differences from the strongest reported fixed baseline for each setting in Tables 2 and 3. Positive values favor adaptive fusion. The panels use their respective evaluation metrics and axis ranges; all target settings are shown.