WAVEHUMANIZE: LOCALIZED TEXT HUMANIZATION WITH WAVELET ENERGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Text humanization aims to make LLM-generated writing more human-like while preserving its meaning. Achieving this with minimal changes to the original text requires identifying which passages to revise and choosing suitable replacements. We propose **WaveHumanize**, a training-free framework that uses a shared wavelet-energy signal to guide both decisions. A pretrained language model scores the tokens in a document based on its next-token predictions. Wavelet analysis converts these scores into an energy profile aligned with text positions. We use this profile to select short spans for rewriting, and a second language model generates candidate replacements. Each candidate is inserted into the document and rescored, and the replacement with the highest average energy over its span is retained. The procedure requires no feedback from the target detector. In the primary evaluation across fifteen domains, WaveHumanize achieves greater detection reduction with less textual change than the compared limited-edit baselines. At thresholds calibrated to a 1% false-positive rate, an average of 80.8% of initially detected texts evade detection after rewriting. Further evaluation shows transfer to several additional detector families. Code is available at `https://anonymous.4open.science/r/WaveHumanize-ICLR-C6FE`.

1 INTRODUCTION

Large Language Models (LLMs) have become increasingly useful writing assistants, helping users develop ideas and turn them into drafts. However, fluent machine-generated text does not always match the style that writers seek. Compared with human writing, LLM-generated text can exhibit noticeable differences in wording, lexical diversity, and expression. This creates a need to revise generated text even when its content is already suitable. Text humanization addresses this problem by rewriting LLM-generated text to make it more human-like while preserving its meaning (Krishna et al., 2023; Huang et al., 2025).

We focus on humanization through a small number of targeted edits, retaining most of the original text. Existing approaches use specialized paraphrasers (Krishna et al., 2023), simulated high-temperature generation (Huang et al., 2025), and detector-guided optimization (Cheng et al., 2025). Under a limited edit budget, however, it is also important to decide where these edits should be applied. Selective methods such as RAFT use proxy scores to identify words for revision, while relying on target-detector feedback to select replacements (Wang et al., 2024a). We seek a statistical signal that can guide both edit placement and replacement selection without querying the target detector.

Statistical text detection provides a useful starting point. Given a document, a pretrained language model can assign a probability to each observed token based on the preceding context. These probabilities form a sequence along the document. Existing detection methods summarize such sequences using statistics such as likelihood or log-rank (Su et al., 2023), while others analyze their spectral structure (Xu et al., 2024). Sentence-level detection further shows that these sequences contain information about where generated passages occur (Wang et al., 2023). These findings suggest that statistical patterns can also provide guidance for selective humanization. If such patterns are kept aligned with text positions, they can be used to identify candidate regions for editing and to compare possible rewrites.

To address this challenge, we propose **WaveHumanize**, a training-free framework that uses wavelet energy to guide both edit placement and replacement selection. A pretrained language model first

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.

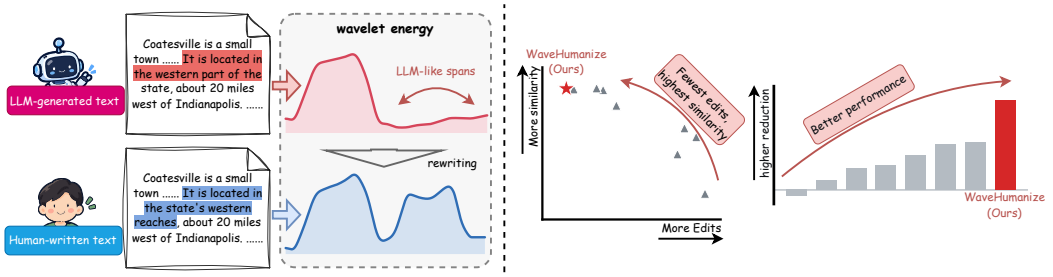

Figure 1: WaveHumanize overview and empirical comparisons. **Left:** an example of localized rewriting with the corresponding wavelet-energy profiles before and after editing. Highlighted text shows the original span and its replacement. **Right:** the scatter plot compares measured edit cost and source-text similarity, with WaveHumanize shown as a red star and other methods as grey triangles. The bar plot compares detectability reduction, with WaveHumanize highlighted in red.

assigns a log-rank score to each observed token. We then apply a Haar wavelet approximation to the resulting sequence and construct an energy profile over the document. This profile summarizes neighboring deviations from the document-level average while retaining their positions in the text. In diagnostic experiments containing both human-written and machine-generated passages, lower-energy regions are more likely to correspond to generated text, motivating their use as candidate edit regions.

WaveHumanize first excludes high-energy regions and samples short, non-overlapping spans from the remaining text. An instruction-tuned language model generates several replacements for each selected span. Each candidate is then reinserted into the document and rescored in context, and the candidate with the highest average energy over the replacement span is retained. The profile is recomputed before the next editing round. In this way, the same statistical signal guides both where to edit and which rewrite to choose, without additional training or target-detector queries. Figure 1 provides a compact overview of this process and its empirical behavior: the left panel illustrates localized rewriting together with the corresponding wavelet-energy profiles before and after editing, while the right panels compare measured edit cost, source-text similarity, and detectability reduction across methods.

We evaluate WaveHumanize across fifteen domains from MAGE and HC3 (Li et al., 2024; Guo et al., 2023). Results show that WaveHumanize reduces detectability with substantially less textual change than the compared baselines, while its gains transfer to several additional detector families. Our main contributions are:

- We introduce WaveHumanize, a training-free framework that uses a shared wavelet-energy signal to guide both edit placement and replacement selection.
- We show that localized statistical guidance enables stronger humanization under a limited edit budget, achieving better detection reduction with less textual change than the evaluated baselines.
- We provide controlled analyses showing that both guided span selection and energy-based candidate ranking contribute to the improvement, with transfer across multiple detector families.

2 Related Work

2.1 Machine-Generated Text Detection

Post-hoc detectors use supervised classification (Ippolito et al., 2020), including adversarial training for paraphrase robustness (Hu et al., 2023), or training-free statistics: likelihood and log-rank (Su et al., 2023), probability curvature (Mitchell et al., 2023; Bao et al., 2024), and cross-model comparisons (Hans et al., 2024). Watermarking instead embeds signals during generation (Kirchenbauer et al., 2023). Broad benchmarks (Dugan et al., 2024; Wang et al., 2024b) and perturbation studies (Huang

et al., 2024; Sadasivan et al., 2024) expose generalization and robustness challenges. We study robustness under local rewriting, separating the proxy that guides edits from the detectors used for evaluation.

2.2 Local and Spectral Representations

Localization research examines human recognition of human–machine text boundaries (Dugan et al., 2023), token-level inspection (Gehrmann et al., 2019), and sentence-level detection (Wang et al., 2023). Complementary spectral approaches characterize relative-likelihood sequences (Xu et al., 2024), Fourier energy (Luo et al., 2026), and learned wavelet representations (Liu et al., 2026). These approaches support recognition and detection; rewriting additionally requires deciding where and how to intervene, and what none of them provides is a position-resolved signal used to select the intervention itself. WaveHumanize uses a fixed Haar approximation-energy profile to guide both span selection and replacement ranking.

2.3 Humanization and Selective Rewriting

Humanization rewrites generated text to reduce detectability; TH-Bench evaluates its effectiveness, text quality, and computational cost (Zheng et al., 2025). Approaches include trained paraphrasing (Krishna et al., 2023), simulated high-temperature generation (Huang et al., 2025), style adaptation (Gu et al., 2026), and detector-guided optimization (Cheng et al., 2025). Local attacks use small perturbations (Zhou et al., 2024); RAFT selects substitutions using proxy-based word selection together with target-detector queries (Wang et al., 2024a). Thus, selective editing already has precedents. The constraint that distinguishes our setting is that the evaluated detector is the object of study, so it cannot be queried during rewriting, and guidance must come from the proxy alone. WaveHumanize addresses this constraint through one proxy-derived energy profile that selects lower-energy spans and ranks replacements after contextual reinsertion, without additional training.

3 Method

3.1 Overview

WaveHumanize uses a shared local energy profile to decide *where to edit* and *which replacement to retain*. A frozen proxy language model converts the document into token-level log-rank scores, from which we construct a Haar approximation-energy profile. We select non-overlapping spans from the lower-energy portion of the document, generate contextual replacements with an instruction-tuned model, and retain the candidate with the highest mean local energy after reinsertion. Both models remain frozen, and the evaluated detector is unavailable throughout rewriting. Figure 2 summarizes the procedure.

3.2 Problem Setup

Given a machine-generated document x, we seek a rewrite x' that reduces detectability through limited local changes while preserving meaning. Let $d(\cdot)$ be an evaluation detector whose score increases with evidence of machine generation. The desired reduction is

$$\Delta_d(x, x') = d(x) - d(x'). \tag{1}$$

Under our protocol, d cannot be queried during rewriting. We therefore obtain editing guidance from a frozen proxy language model p_θ and generate replacements with a frozen instruction-tuned model q_ϕ.

The procedure controls intervention size through a per-round budget on selected source tokens. Realized textual change is measured separately by token-sequence dissimilarity $\tau(x, x')$, as defined in Appendix E. Meaning preservation is encouraged through contextual generation instructions and assessed after rewriting; the deployed procedure does not enforce a semantic-similarity threshold.

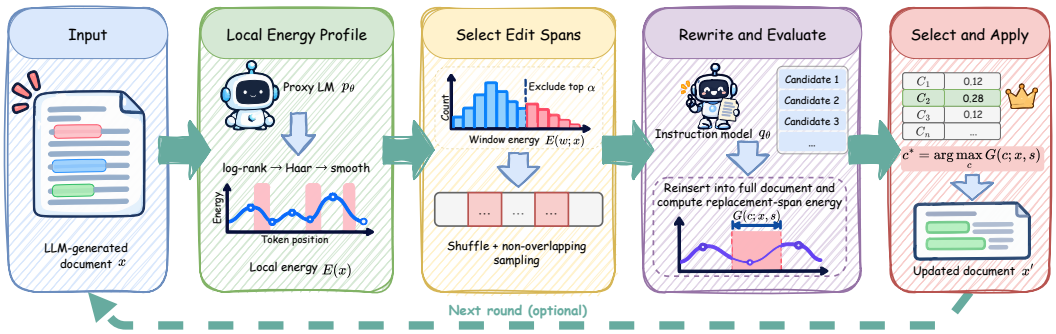

Figure 2: Overview of WaveHumanize. A frozen proxy model provides token log-ranks, which are converted into a local Haar approximation-energy profile. High-energy windows are excluded, and non-overlapping edit spans are selected from the remaining pool. An instruction-tuned model proposes replacements, each of which is reinserted into the document and rescored. The replacement with the highest mean energy over its own span is retained.

3.3 Local Energy Profile

Editing decisions concern spans, so the guidance signal should summarize neighboring token statistics while remaining aligned with edit locations. We construct this signal from proxy log-ranks using a fixed, local Haar transform.

Let $(x_0, \ldots, x_{T-1})$ be the proxy-tokenized sequence within its context limit. For each predicted token, we compute

$$r_i = \log \operatorname{rank}_{p_\theta}\left(x_{i+1} \mid x_{0:i}\right), \qquad 0 \le i < N, \quad N = T - 1, \tag{2}$$

where rank is one-indexed in descending predictive probability. Larger r_i place the observed token farther down the proxy's probability ranking for its context. We remove the document-level mean:

$$z_i = r_i - \bar{r}, \qquad \bar{r} = \frac{1}{N} \sum_{j=0}^{N-1} r_j. \tag{3}$$

We summarize adjacent scores using the level-one Haar approximation (Mallat, 1989; Daubechies, 1992):

$$a_k = \frac{z_{2k} + z_{2k+1}}{\sqrt{2}}, \qquad 0 \le k < L, \quad L = \lceil N/2 \rceil. \tag{4}$$

For odd N, the final value is repeated by symmetric extension. This transform retains positional information while aggregating adjacent deviations from the document mean. Its squared coefficients measure the magnitude of these local averages: for example, centered pairs $(2, 2)$ and $(2, -2)$ yield approximation energies 8 and 0, respectively, before smoothing.

We smooth the squared coefficients over a local neighborhood:

$$e_k = \frac{1}{|J_k|} \sum_{j \in J_k} a_j^2, \qquad J_k = \{j : 0 \le j < L,\ |j - k| \le h\}, \tag{5}$$

with $h = 2$, corresponding to a smoothing width of five coefficients. The denominator accounts for boundary truncation. Linear interpolation from coefficient positions k to token positions $i/2$ gives the profile $E_i(x)$, with endpoint values extended when needed.

The profile measures local departures in the centered proxy statistic. It is an editing signal, not a calibrated probability of machine authorship. The deployed representation uses only the level-one approximation component; alternative representations are examined in Appendix K.

3.4 Edit Region Selection

We use the profile to allocate a limited number of edits to lower-energy regions. After sentence segmentation and noise filtering, candidate windows contain $W = 5$ consecutive scored tokens within

a sentence and are enumerated with stride one. Sentences with fewer than $\max(2, W)$ scored tokens contribute no windows. For a candidate window w with token positions I_w, its energy is

$$E(w; x) = \frac{1}{W} \sum_{i \in I_w} E_i(x). \tag{6}$$

Let $\mathcal{W}(x)$ denote the candidate-window set. We retain its lower-energy portion using an empirical quantile:

$$\begin{aligned} t_\alpha(x) &= Q_{1-\alpha}\big(\{E(w; x) : w \in \mathcal{W}(x)\}\big), \\ \mathcal{P}(x) &= \{w \in \mathcal{W}(x) : E(w; x) \le t_\alpha(x)\}. \end{aligned} \tag{7}$$

We set $\alpha = 0.30$, retaining windows at or below the empirical 0.70 quantile, including ties.

This **High-Energy Exclusion** rule screens candidate locations without assuming that progressively lower energy implies progressively greater editing benefit. We therefore retain a broader lower-energy pool rather than concentrate edits at its minima. Comparisons with Lowest-Energy and Random localization in Section 4 test this design choice.

We randomly permute $\mathcal{P}(x)$ using a document-specific random stream and greedily accept token-disjoint windows until the pool is exhausted or the per-round budget is reached:

$$B = W \max\{1, \mathrm{round}(\rho m)\}, \tag{8}$$

where m counts sentences remaining after noise filtering and $\rho = 0.30$ controls the nominal selection budget. Documents with fewer than two such sentences are left unchanged. The random stream is retained across rounds.

Overlap constraints or a small retained pool can leave the budget partially unused. Moreover, B counts selected source tokens in one round; it does not directly bound the final sequence dissimilarity $\tau(x, x')$. Preprocessing and randomization details appear in Appendix A.

3.5 Candidate Generation and Selection

For each selected character span $s = [a, b)$, q_ϕ samples up to $K = 20$ replacements conditioned on the fragment and its containing sentence, using temperature 0.9 and top-$p = 0.95$. The prompt requests changes only to the selected fragment while preserving meaning, facts, tone, and approximate length. We remove empty, duplicate, unchanged, clearly truncated, and excessively long outputs, yielding the candidate set $\mathcal{C}_s$. The prompt, generation-length limits, and filtering details are provided in Appendices A and C.

Candidate quality is evaluated in the document context. For each $c \in \mathcal{C}_s$, we restore the original fragment's boundary whitespace to obtain $\widetilde{c}$ and construct

$$x^{(c)} = x[:a] \| \widetilde{c} \| x[b:],$$

where $\|$ denotes concatenation. We then retokenize $x^{(c)}$ and recompute its proxy scores, document mean, and energy profile using the same context limit. This accounts for the candidate's effect on contextual predictions and normalization.

Let I_c contain the rescored token positions whose character starts fall within the replacement interval. We score and select candidates by

$$G(c; x, s) = \frac{1}{|I_c|} \sum_{i \in I_c} E_i(x^{(c)}), \qquad c^* = \arg\max_{c \in \mathcal{C}_s} G(c; x, s). \tag{9}$$

If I_c is empty, its score is set to zero. Averaging over replacement tokens permits comparison between candidates of different lengths. The rule favors greater local approximation energy among valid rewrites; it does not use the evaluation detector.

A span remains unchanged when $\mathcal{C}_s$ is empty. Otherwise, the highest-scoring candidate is retained without requiring an improvement over the original span, which is not included in the candidate set. All candidates within a round are evaluated against the same current document. Selected replacements are then applied from right to left to preserve character offsets.

We recompute the profile and edit locations on the updated document for at most $R = 2$ rounds, stopping early if no replacement is applied. Thus, localization and ranking share one statistic while operating at different stages: localization selects lower-energy source regions, and ranking compares their contextual replacements. Algorithm 1 in Appendix M summarizes the procedure.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate whether WaveHumanize reduces detectability at low textual edit cost. Component analyses examine the contributions of local representation, span selection, and candidate ranking. We then test transfer across detectors and performance at fixed thresholds.

Data. We evaluate 49,043 machine-generated texts and 68,186 human-written references from fifteen domains in MAGE (Li et al., 2024) and HC3 (Guo et al., 2023). For compact notation in figures, we refer to MAGE as Corpus A and HC3 as Corpus B. Texts contain 30–400 words.

Models. GPT-J-6B provides the proxy token statistics, and Meta-Llama-3-8B-Instruct generates candidate replacements. An independent GPT-Neo-2.7B likelihood detector serves as the primary evaluator. The evaluator is never queried during rewriting. Full settings appear in Appendix A.

Metrics. We measure detectability reduction as $\Delta A = A^{\text{orig}} - A^{\text{rew}}$, where A denotes detection AUROC. Positive values indicate reduced detectability. We measure textual edit cost using token-sequence dissimilarity $\overline{\tau}$ and report embedding similarity as a proxy for meaning preservation. Within each domain, textual-edit efficiency is $\eta = \Delta A/\overline{\tau}$. Macro efficiency is the mean of these domain-level ratios. AUROC summarizes ranking across thresholds, while fixed-threshold evaluation measures whether individual texts cross a calibrated decision boundary. Full definitions appear in Appendix E.

Baselines and budget. We compare three external method families and two unguided controls. TempParaphraser (TP) uses either one candidate ($N = 1$) or seven candidates selected by its detection oracle ($N = 7$) (Huang et al., 2025). Adversarial Paraphrasing (AP) is evaluated with random sentence selection and with RADAR-guided selection (Cheng et al., 2025; Hu et al., 2023). DIPPER paraphrases the selected sentences (Krishna et al., 2023). Simple and Recursive use the AP implementation with adversarial guidance disabled, rewriting each selected sentence once or twice.

All methods receive a nominal local edit budget of 30%. External methods retain their own sentence-selection rules, while WaveHumanize selects five-token spans, so realized textual change and computational cost can differ. We report $\overline{\tau}$ for every method and score all outputs with the same independent evaluator. The shared nominal budget thus fixes the intended scope of intervention, while measured dissimilarity captures how much each method actually changes the text. Baseline configurations appear in Appendix F.

4.2 HUMANIZATION PERFORMANCE AND EDIT COST

Stronger reduction at lower textual edit cost. WaveHumanize achieves the largest AUROC reduction among the compared local-budget methods in all fifteen domains (Table 1). Its macro ΔA is 0.224, compared with 0.119 for AP, the strongest external baseline by this measure. Macro token-sequence dissimilarity is 0.116, the lowest among the compared methods, and embedding similarity is 0.968. Macro textual-edit efficiency reaches 1.881, or $3.9\times$ that of TP ($N = 1$), the strongest external baseline by efficiency. Figure 3(a) summarizes the trade-off between detectability reduction and textual change; its efficiency rays use ratios of macro means rather than the mean of domain-level ratios reported above.

The unguided controls produce smaller reductions despite higher edit costs: Simple and Recursive reach macro ΔA values of 0.024 and 0.055, with dissimilarities of 0.261 and 0.297. DIPPER yields a negative macro reduction (-0.015) under the local-budget protocol. More extensive rewriting therefore does not necessarily produce a larger reduction in detectability. For comparison, full-document rewriting produces token-sequence dissimilarities of 0.536–0.693 for TP, 0.682–0.839 for

	Unguided		TempParaphraser		Adversarial para.			
Domain	Simple	Recursive	$N{=}1$	$N{=}7$	random	RADAR	DIPPER	WaveHumanize
CMV	-0.018	0.013	0.104	0.060	0.129	0.125	-0.052	**0.293**
CS/AI	0.070	0.120	0.077	0.078	0.177	0.172	0.012	**0.205**
ELI5	-0.004	0.031	0.089	0.055	0.124	0.130	-0.038	**0.278**
Finance	0.004	0.007	0.013	0.012	0.021	0.014	0.001	**0.023**
HellaSwag	-0.018	0.010	0.097	0.050	0.097	0.076	-0.032	**0.290**
Medicine	0.001	0.002	0.004	0.004	0.006	0.005	0.000	**0.023**
Open QA	0.093	0.134	0.104	0.102	0.181	0.170	0.028	**0.207**
Reddit ELI5	0.004	0.009	0.011	0.011	0.021	0.021	0.002	**0.041**
ROCStories	0.101	0.155	0.155	0.119	0.230	0.231	-0.006	**0.442**
SCI	0.047	0.104	0.109	0.066	0.161	0.143	-0.016	**0.240**
SQuAD	0.013	0.046	0.102	0.069	0.126	0.116	-0.024	**0.219**
TLDR	0.026	0.069	0.143	0.113	0.154	0.152	-0.006	**0.279**
WritingPrompts	-0.001	0.023	0.084	0.048	0.102	0.099	-0.031	**0.279**
XSum	0.048	0.081	0.112	0.090	0.156	0.140	-0.010	**0.225**
Yelp	-0.002	0.021	0.098	0.052	0.105	0.112	-0.054	**0.318**
Macro $\Delta A \uparrow$	0.024	0.055	0.087	0.062	0.119	0.114	-0.015	**0.224**
Macro $\overline{\tau} \downarrow$	0.261	0.297	0.189	0.160	0.286	0.272	0.140	**0.116**
Macro Sim. $\uparrow$	0.939	0.918	0.969	**0.973**	0.926	0.927	0.966	0.968
Macro $\eta \uparrow$	0.099	0.190	0.477	0.424	0.417	0.418	-0.094	**1.881**

Table 1: Main results across fifteen domains under the independent GPT-Neo-2.7B likelihood evaluator. Each cell is the AUROC reduction ΔA of one configuration in one domain, as defined in Section 4.1; bold marks the best value in a row and underline the second best. The last block gives unweighted macro averages over the fifteen domains of ΔA, token-sequence dissimilarity $\overline{\tau}$, embedding similarity, and textual-edit efficiency $\eta = \Delta A/\overline{\tau}$. Every configuration operates under the same nominal local edit budget, and realized edit costs are measured separately. WaveHumanize is placed after the vertical rule because it is the method proposed here rather than a baseline; Section 4.1 describes each baseline.

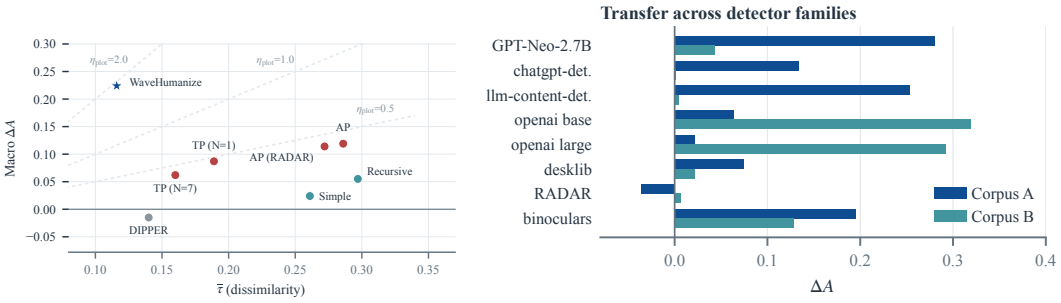

Figure 3: Humanization performance, textual edit cost, and detector transfer. **(a)** Macro AUROC reduction versus macro token-sequence dissimilarity. Dashed rays, labelled η_{plot} in the panel, represent equal ratios of the plotted macro means. Table 1 instead reports the mean of domain-level efficiency ratios. **(b)** Transfer of frozen rewritten outputs to additional detectors, with results pooled separately within each corpus. Positive values indicate reduced detectability.

AP, and 0.440–0.534 for DIPPER across domains, substantially above the macro value of 0.116 for WaveHumanize. Full-document baseline results appear in Appendix G.

Variation across domains. AUROC reduction reaches 0.442 on ROCStories, 0.318 on Yelp, and 0.293 on CMV, and is smaller on Finance and Medicine, at 0.023 each. The evaluator strongly separates human and machine text in these two domains before rewriting, so much of that separation remains. The magnitude of the benefit therefore varies substantially across domains.

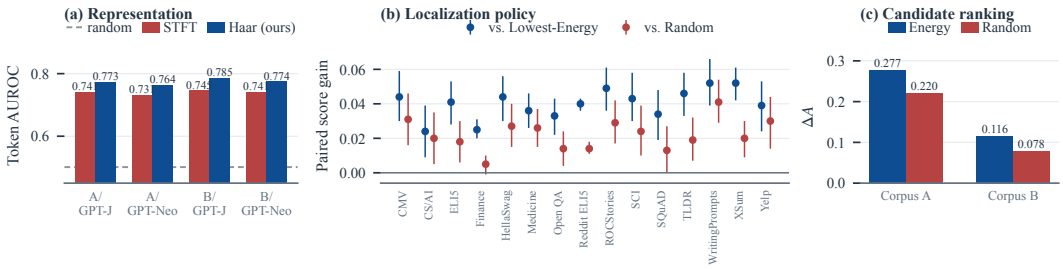

Figure 4: Component analysis. **(a)** Controlled span-localization diagnostic comparing Haar approximation energy with matched-window STFT across two corpora and two proxy models. **(b)** Full-data comparison of High-Energy Exclusion against Lowest-Energy and Random localization. Whiskers show 95% paired document-bootstrap intervals for evaluator-score differences. **(c)** Energy-based versus random candidate ranking on the diagnostic subset, with localization, generation, and candidate pools held fixed.

4.3 Component Analysis

We first assess whether the representation provides useful local information. We then test how span selection and candidate ranking affect rewriting. The localization-policy comparison uses the full evaluation data. The representation and candidate-ranking studies use separate diagnostic data.

Local representation. We test the representation without invoking the rewriter, using documents that combine a human-written prefix with a machine-generated continuation. Negated local energy serves as the score for identifying generated tokens. Across four corpus–proxy settings, Haar approximation energy improves average token AUROC from 0.739 for matched-window STFT to 0.774. AUPRC increases from 0.742 to 0.758 (Figure 4a).

The advantage persists after removing boundary-adjacent continuation tokens. Haar obtains AUROC/AUPRC of $0.777/0.751$, compared with $0.742/0.734$ for STFT. A random score profile yields AUROC 0.501. These results support the profile as a local signal beyond the immediate splice boundary. This diagnostic isolates localization quality before the following ablations test its usefulness for rewriting. Full representation results appear in Appendix K.

Span selection. We hold the rewriter, fragment width, candidate count, and round limit fixed, and compare the deployed High-Energy Exclusion policy against lowest-energy and random selection from the same windows.

On the full evaluation data, the deployed policy produces larger decreases in the evaluator's likelihood score than Lowest-Energy selection in all fifteen domains. Paired score advantages range from 0.024 to 0.052. All domain-wise 95% intervals exclude zero, with no adjustment for multiple comparisons.

Point estimates also favor the deployed policy over Random selection in all fifteen domains, although the intervals overlap zero on Finance and SQuAD. At the macro level, the paired score advantage is 0.040 over Lowest-Energy with a corresponding dissimilarity difference of -0.0012, and 0.023 over Random with a difference of 0.004. A separate diagnostic comparison finds Highest-Energy selection weaker on both corpora. Per-domain results appear in Appendix K.

Candidate ranking. On a domain-balanced diagnostic subset of 2,000 text pairs, we hold localization, generation, candidate pools, and the round limit fixed. Energy-based ranking increases macro ΔA from 0.174 under uniform-random ranking to 0.226, a gain of 0.052 (Table 15). Improvements occur in all fifteen domains. Per-domain differences in dissimilarity and similarity are at most 0.009 and 0.008. Because both variants choose from the same generated candidates, this comparison isolates the ranking decision rather than candidate generation. Figure 4(c) presents the corresponding corpus-pooled results.

Alternative localization statistics. Separate diagnostic runs replace the localization statistic with alternative wavelet bands, their multiscale combination, or sliding-window variance, leaving candidate

ranking unchanged. Across the alternative wavelet representations, ΔA ranges from 0.244 to 0.262 on MAGE and from 0.127 to 0.130 on HC3, compared with 0.210 and 0.094 under random localization. AUROC reductions persist on both corpora (Appendix K), so the effect is not confined to the deployed wavelet band, although these comparisons do not establish an advantage for multiscale localization.

4.4 Transfer Across Detectors

We apply additional detectors to the frozen rewritten outputs, all outside the rewriting loop. Figure 3(b) reports corpus-pooled results.

Detectability decreases under several detector families. Binoculars yields AUROC reductions of 0.195 on MAGE and 0.128 on HC3. The two OpenAI-detector reproductions show larger reductions on HC3, at 0.319 and 0.292. RADAR is an exception: ΔA is -0.036 on MAGE and 0.007 on HC3. Transfer therefore extends beyond the primary likelihood evaluator without direct detector access during rewriting, although its strength and direction depend on the detector; per-domain results appear in Appendix J.

4.5 Fixed-Threshold Validation

We calibrate a threshold in each domain using half of its human reference texts, targeting a 1% false-positive rate; the remaining texts estimate the held-out rate, which ranges from 0.2% to 2.9%. The threshold remains unchanged for original and rewritten machine texts.

WaveHumanize achieves a macro conditional attack success rate (ASR) of 80.8%. Within each domain, conditional ASR is the fraction of initially detected texts that evade detection after rewriting. The macro result averages these domain-level rates. It therefore describes changes among texts the detector originally identified, rather than the fraction of all machine texts labeled human after rewriting. Performance ranges from 22.0% on Medicine to 98.9% on CS/AI, showing substantial variation at this operating point.

Adding a lower-tail criterion at the human 0.5% calibration quantile (nominal total 1.5% false-positive rate) recaptures only 11 of the $15{,}651$ one-sided escapes (0.07%), so opposite-tail overshoot accounts for few escapes. Per-domain operating points and the two-sided evaluation appear in Appendix L.

5 Discussion

Local statistical guidance extends spectral detection (Xu et al., 2024; Liu et al., 2026) to selective rewriting. Sampling from a broader lower-energy pool outperformed minimum-energy selection, suggesting that lower energy does not guarantee greater editing benefit. Energy therefore serves as a screening signal. Reductions with alternative estimators show that useful localization is not confined to a single wavelet representation. Transfer remained detector-dependent, with negative transfer under RADAR on MAGE. One proxy–rewriter pair limits cross-model generalization. Fixed detectors leave adaptive robustness untested, while automatic quality measures do not establish perceived naturalness.

6 Conclusion

This work investigates selective local intervention for text humanization without target-detector feedback during rewriting. We introduce WaveHumanize, a training-free framework that converts proxy-model token statistics into a position-resolved wavelet-energy profile. This shared signal guides edit-region selection and contextual candidate ranking. Across fifteen domains, WaveHumanize reduces detectability with substantially less textual change than the compared local-budget baselines. At thresholds calibrated to a 1% false-positive rate, conditional evasion among initially detected texts averages 80.8% across domains. Macro textual-edit efficiency is $3.9\times$ that of TempParaphraser ($N = 1$), the strongest external baseline by this metric. Controlled analyses support both editing decisions: where to edit and which replacement to retain. Independent evaluations further show transfer to several detector families. Together, these findings support local statistical guidance as a practical way to allocate a limited rewriting budget.

AI USE STATEMENT

Generative AI assistance was used to inspect the supplied code and experiment records, organize the argument, draft and edit the manuscript, assemble references, and prepare the document-generation scripts. The rewriting methods under study also use generative language models, as described in Section 3. Those are distinct roles. The experimental results reported here were produced by the described rewriting pipeline and its scoring scripts; generative assistance was confined to the writing of the manuscript and to the tooling that assembles it.

ETHICS STATEMENT

Localized rewriting can support controlled editing and analysis of detector sensitivity, and it can also be misused to conceal machine-generated authorship or to weaken provenance mechanisms. A detector-score change therefore should not be read as evidence of human authorship, of factual correctness, or of the legitimacy of a text. We recommend that generated or substantially rewritten content continues to be disclosed as such. No new human-subject study was conducted for this version.

REPRODUCIBILITY STATEMENT

Section 3 specifies the local statistical signal, the High-Energy Exclusion policy, candidate ranking, and the iteration schedule. The appendices record the remaining details: the complete configuration and preprocessing pipeline in Appendix A, the scoring rules, calibration protocol, and data alignment in Appendix E, the configuration of every compared system in Appendix F, and the controlled diagnostics in Appendix K. Per-domain, per-detector, and fixed-threshold results appear in Appendices G–L. An anonymized code package accompanies this submission. It contains the aggregation values behind the reported numbers, the scripts that generate the tables and figures, the processing steps for MAGE and HC3, and a source manifest pinning the inputs used for each reported number.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 24814–24836, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6b8c6f846c3575e1d1ad496abea28826-Paper-Conference.pdf.

Yize Cheng, Vinu Sankar Sadasivan, Mehrdad Saberi, Shoumik Saha, and Soheil Feizi. Adversarial paraphrasing: A universal attack for humanizing AI-generated text. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=fYjF9KIJd5.

Ingrid Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992. doi: 10.1137/1.9781611970104. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611970104.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. Real or fake text? investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12763–12771, 2023.

Liam Dugan, Alyssa Hwang, Filip Trhlik, Josh Magnus Ludan, Andrew Zhu, 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)*. Association for Computational Linguistics, 2024.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In Marta R. Costa-jussà and Enrique Alfonseca (eds.), *Proceedings of*

the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 111–116, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3019. URL https://aclanthology.org/P19-3019/.

Yongtong Gu, Songze Li, and Xia Hu. MASH: Evading black-box AI-generated text detectors via style humanization. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29749–29769, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1487. URL https://aclanthology.org/2026.findings-acl.1487/.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is ChatGPT to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023. URL https://arxiv.org/abs/2301.07597.

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*, ICML'24. JMLR.org, 2024.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning, 2023. URL https://arxiv.org/abs/2307.03838.

Guanhua Huang, Yuchen Zhang, Zhe Li, Yongjian You, Mingze Wang, and Zhouwang Yang. Are AI-generated text detectors robust to adversarial perturbations? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6005–6024, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.327. URL https://aclanthology.org/2024.acl-long.327/.

Junjie Huang, Ruiquan Zhang, Jinsong Su, and Yidong Chen. TempParaphraser: "heating up" text to evade AI-text detection through paraphrasing. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31554–31573, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1607. URL https://aclanthology.org/2025.emnlp-main.1607/.

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

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. PMLR, 2023.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense, 2023. URL https://arxiv.org/abs/2303.13408.

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, 2024. doi: 10.18653/v1/2024.acl-long.3. URL https://aclanthology.org/2024.acl-long.3/.

Zhichen Liu, Kaitong Qin, Linhan He, and Yang Xu. WaveDetect: Robust framework for machine-generated text detection via wavelet transform. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics:*

ACL 2026, pp. 8712–8727, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.424. URL https://aclanthology.org/2026.findings-acl.424/.

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. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i38.40510. URL https://doi.org/10.1609/aaai.v40i38.40510.

S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(7):674–693, July 1989. ISSN 0162-8828. doi: 10.1109/34.192463. URL https://doi.org/10.1109/34.192463.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24950–24962. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/mitchell23a.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-generated text be reliably detected? *Transactions on Machine Learning Research*, 2024.

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

James Liyuan Wang, Ran Li, Junfeng Yang, and Chengzhi Mao. RAFT: Realistic attacks to fool text detectors. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16923–16936, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.939. URL https://aclanthology.org/2024.emnlp-main.939/.

Pengyu Wang, Lili Li, Kezhi Ren, Bo Jiang, Dong Zhang, and Xipeng Qiu. SeqXGPT: Sentence-level AI-generated text detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Chenxi Whitehouse, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. M4GT-Bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2024b.

Yang Xu, Yu Wang, Hao An, Zhichen Liu, and Yongyuan Li. Detecting subtle differences between human and model languages using spectrum of relative likelihood. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10108–10121, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.564. URL https://aclanthology.org/2024.emnlp-main.564/.

Jingyi Zheng, Junfeng Wang, Zhen Sun, Wenhan Dong, Yule Liu, and Xinlei He. Th-bench: Evaluating evading attacks via humanizing ai text on machine-generated text detectors. In

Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD '25, pp. 5948–5959, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3737418. URL https://doi.org/10.1145/3711896.3737418.

Ying Zhou, Ben He, and Le Sun. Humanizing machine-generated content: Evading AI-text detection through adversarial attack. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 8427–8437, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.739/.

Item	Recorded configuration
Proxy	GPT-J-6B, FP16
Rewriter	Meta-Llama-3-8B-Instruct, FP16
Local statistic	Level-one Haar approximation energy
Coefficient smoothing	Width 5, boundary-normalized mean
Localization	Exclude windows above the 0.70 energy quantile; shuffle the retained pool and greedily select token-disjoint windows up to budget B
Budget	$B = W \max\{1, \mathrm{round}(\rho m)\}$, with $\rho = 0.30$
Window	$W = 5$ proxy tokens, constrained to a sentence span
Candidates / rounds	20 candidate draws per selected span / at most 2 rounds
Generation	Temperature 0.9, top-$p = 0.95$; sampling enabled
Max new tokens	$\min(200, \max(40, \ell_{\max} + 60))$ for fragment rewriting
Similarity filter	Disabled; similarity is used only for evaluation
Random seeds	Generation: 42; localization: 42 plus document input index, with the stream retained across rounds
Similarity model	`all-MiniLM-L6-v2`
Environment	Python 3.8, PyTorch 2.1.0+cu121, Transformers 4.44.1
Hardware	NVIDIA H800 80 GB; runs distributed over 4–6 GPUs

Table 2: Configuration used for the full-data experiments. $\ell_{\max}$ denotes the maximum fragment length under the rewriter tokenizer within a generation batch.

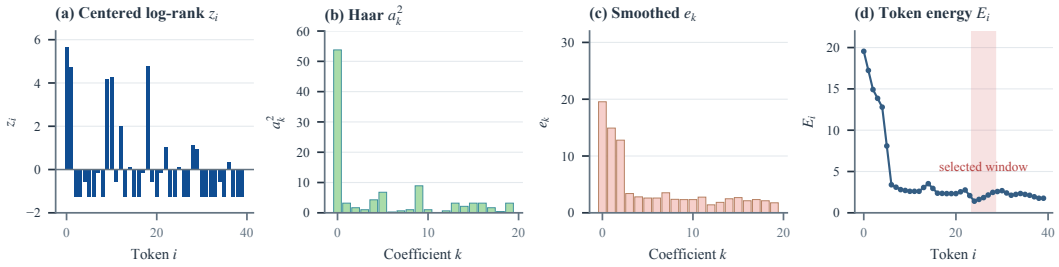

Figure 5: From the centered proxy statistic to the local energy profile on a real 40-token example. **(a)** Centered proxy log-rank sequence z_i. **(b)** Squared level-one Haar approximation coefficients a_k^2, each aggregating one adjacent token pair. **(c)** Smoothed coefficient energy e_k, obtained by averaging a_k^2 over five neighboring coefficient positions. **(d)** Token-level energy E_i, obtained by interpolating e_k back onto the token grid. The shaded five-token interval illustrates a low-energy candidate window; the deployed High-Energy Exclusion rule does not deterministically select the minimum-energy window.

A Configuration and Preprocessing

Documents are segmented with a punctuation/newline regular expression followed by noise-span filtering. A span is treated as noise if, after stripping, it contains fewer than two characters or contains no Latin letter, digit, or CJK character. The remaining sentence count determines the per-round editing budget in Equation (8); sentence boundaries also constrain candidate windows and provide generation context. Token membership is determined from character start offsets.

The implementation uses Python's `round`, including ties-to-even behavior. Documents are filtered using inclusive 30–400 word bounds. Human and machine pools are retained separately after filtering and are not forced to have equal sizes within a domain; all reported domain-level metrics use the corresponding frozen human and machine pools.

B Localization Details

Candidate windows are enumerated with stride one within every non-noise sentence containing at least W scored tokens; windows never cross sentence boundaries. High-Energy Exclusion computes the linearly interpolated 0.70 quantile of the current window-energy distribution and retains all

windows at or below that threshold. The retained pool is shuffled using the document-specific NumPy `RandomState`, after which windows are greedily accepted subject to token non-overlap and duplicate-character-interval constraints.

Selection terminates when the token budget B is reached or when no eligible windows remain. Consequently, overlaps or a small retained pool may leave the nominal budget partially unfilled. No content-word or minimum-character heuristic is applied by the deployed level-one localization rule. Figure 5 illustrates the construction of the local energy profile rather than a deterministic localization decision.

C Generation Prompt and Candidate Handling

WaveHumanize prompt. The system prompt used for fragment rewriting is:

> You are an expert editor. Rewrite the highlighted part of the sentence so it reads more naturally and human-written, with lively rhythm and varied word choice. STRICTLY preserve the original meaning, facts, tone, and roughly the same length. Do not add new information. Output only the rewritten part, nothing else.

The corresponding user message is:

> Sentence: [containing sentence]
>
> Rewrite ONLY the highlighted part to sound more natural and human-written, keeping the rest of the sentence unchanged. Output only the rewritten part.
>
> Part to rewrite: [selected fragment]

Candidates are generated with left padding and decoded only after the prompt boundary. We strip surrounding whitespace, remove empty or duplicate outputs, discard candidates identical to the source fragment, and filter clearly incomplete or excessively long outputs. Because duplicates and invalid generations are removed, fewer than 20 distinct candidates may remain for ranking.

Each surviving candidate is individually reinserted into the full document before rescoring. Retokenization therefore allows the replacement to affect contextual predictions outside the rewritten fragment, although Equation 9 averages the resulting local energy only over tokens aligned with the replacement interval. The original fragment is not automatically retained as a candidate, and no positive-improvement constraint is imposed; the selected candidate can therefore, in principle, decrease the local ranking objective.

D Prompts for Compared Systems

For reproducibility, we reproduce the prompt templates used by the compared rewriting systems.

Adversarial Paraphrasing, Simple, and Recursive.

> You are a rephraser. Given any input text, you are supposed to rephrase the text without changing its meaning and content, while maintaining the text quality. Also, it is important for you to output a rephrased text that has a different style from the input text. You can not just make a few changes to the input text. The input text is given below. Print your rephrased output text between tags `<TAG>` and `</TAG>`.

TempParaphraser.

> Rewrite the following text to sound more natural and human-like. Keep the same information and overall structure, but use more casual language, varied sentence structures, and add subtle personal touches.

DIPPER uses its released checkpoint directly and does not require an additional text prompt.

E Evaluation Protocol and Metrics

Domain alignment. The evaluation unit is a source domain. Human and machine texts are filtered to 30–400 whitespace-separated words. Domain labels are recovered from the original source records through exact text matching, with whitespace normalization for question-answer sources. Each rewritten document is joined to its source document by index and verified against the stored original text.

Original and rewritten AUROCs use the same human reference pool and the same machine-document identities within each domain. Unchanged outputs remain in all denominators. Domain tables are recomputed from the saved rewritten texts, cached detector scores, and per-document similarity records rather than by relabeling results from an earlier rewriting configuration. Human and machine pool sizes are not forced to match. AUROC uses the complete frozen human and machine pools for a domain, while fixed-threshold evaluation uses the calibration/test split described below.

Independent primary evaluator. GPT-Neo-2.7B is used for the domain-level primary evaluation throughout the paper, whereas GPT-J-6B is used only as the proxy for localization and candidate ranking. GPT-Neo is applied offline to frozen outputs and is never queried by the rewriting procedure.

All outputs compared within a detection table are rescored using the same evaluator and input policy. GPT-Neo scoring is capped at 2,048 tokens; edit-cost and embedding-similarity measurements are computed on the saved full documents. The seven GPT-Neo criteria below are alternative scoring rules derived from this checkpoint rather than seven independent detector models. Detector-family transfer is evaluated separately in Appendix I.

GPT-Neo scoring criteria. Let

$$\ell_i = \log p(x_{i+1} \mid x_{0:i}), \qquad r_i = \log \mathrm{rank}_p(x_{i+1} \mid x_{0:i}),$$

and let $p_i(v) = p(v \mid x_{0:i})$.

The likelihood score is

$$D_{\mathrm{LL}} = \frac{1}{N} \sum_i \ell_i,$$

and negative log-rank is

$$D_{\mathrm{LR}} = -\frac{1}{N} \sum_i r_i.$$

LRR is defined as $D_{\mathrm{LL}}/D_{\mathrm{LR}}$. The GLTR-style criterion is the fraction of observed tokens ranked among the ten most probable tokens.

For analytic self-reference Fast-DetectGPT,

$$D_{\mathrm{FD}} = \frac{\sum_i (\ell_i - \mu_i)}{\sqrt{\sum_i \sigma_i^2}},$$

where μ_i and σ_i^2 are respectively the mean and variance of $\log p_i(V)$ under $V \sim p_i$.

For the spectral criterion, we center the token log-likelihood sequence, compute its DFT, and average the squared normalized magnitudes over the first half of the coefficients; this quantity is negated before evaluation. Predictive entropy is likewise negated. All seven criteria are therefore oriented so that larger values indicate more machine-like text.

AUROC, edit cost, and editing efficiency. For criterion D and domain d,

$$\Delta A_D(d) = A_D^{\mathrm{orig}}(d) - A_D^{\mathrm{rew}}(d),$$

so positive values denote reduced detectability after rewriting.

The common edit-cost metric is

$$\tau(x, x') = 1 - \mathrm{SequenceMatcher}(u, v).\,\mathrm{ratio}(), \qquad \eta_D(d) = \frac{\Delta A_D(d)}{\overline{\tau}_d},$$

where u and v are GPT-J token sequences. We use the tokenizer defaults and Python's default `autojunk` behavior. The underlying sequence ratio equals $2M/(|u| + |v|)$ for the matches returned by `SequenceMatcher`. Thus, τ is a normalized sequence dissimilarity and should not be interpreted as the literal fraction of changed source tokens or as a runtime measure.

We first average document-level τ values within each domain. Macro summaries then assign equal weight to all fifteen domains. Macro editing efficiency is the unweighted average of the fifteen within-domain efficiency values.

Fixed operating points. Within each domain, a seed-42 permutation splits human references equally into calibration and held-out test sets; split-size ties follow Python's rounding behavior. The one-sided detector threshold is the 99th percentile of calibration scores, and detection uses the strict rule

$$D(x) > t_d.$$

We report the realized false-positive rate on held-out human texts, original and rewritten TPR, and conditional attack success rate

$$\text{ASR} = \frac{\#\{D(x) > t_d,\ D(x') \le t_d\}}{\#\{D(x) > t_d\}}.$$

The threshold remains unchanged after rewriting. Both numerator and denominator are reported because a nominal 1% calibration target need not produce exactly 1% FPR on held-out human data.

Uncertainty. Localization-policy contrasts use 2,000 paired document-bootstrap resamples within each domain with seed 42, and report percentile 95% confidence intervals for mean likelihood-score differences. Candidate-level analyses use 1,000 document-cluster bootstrap resamples with seed 43 so that multiple candidates originating from the same document remain clustered.

F Baseline Configurations

Local-budget protocol. The nominal 30%-budget baselines select approximately 30% of sentences before applying their respective rewriting procedures. WaveHumanize instead edits five-token fragments using the budget defined in Section 3. Consequently, matching the nominal selection ratio does not imply equal realized token changes or equal computational cost. We therefore report the realized token-sequence dissimilarity $\overline{\tau}$ for every comparison. External baselines use their own localization or sampling rules and never reuse the wavelet-based locator.

TempParaphraser (TP). We reproduce TP using its specialized Llama-3.2-3B paraphraser, temperature 1.2, and repetition penalty 1.05. TP ($N = 1$) draws one rewrite candidate, whereas TP ($N = 7$) draws seven and selects among them using its RoBERTa detection oracle. Full-document and 30%-budget configurations are independent runs. Their saved outputs are rescored using the same GPT-Neo protocol as WaveHumanize.

Adversarial Paraphrasing (AP). The AP reproduction uses RADAR-guided decoding (Cheng et al., 2025; Hu et al., 2023), temperature 0.60, and top-p $= 0.99$. The standard 30%-budget configuration selects sentences randomly, whereas AP (RADAR) uses the baseline's own oracle for sentence selection. Full mode rewrites the complete document. Although AP and WaveHumanize share the Llama-3-8B-Instruct checkpoint in this reproduction, their localization and decoding procedures are different.

Simple and Recursive Paraphrase. Simple and Recursive use the AP implementation with adversarial guidance disabled. Under the 30%-budget setting, Simple rewrites each selected sentence once and Recursive rewrites it twice; all unselected sentences remain unchanged. These are direct local-budget generations rather than fragments reconstructed from full-document rewrites.

Per-domain edit cost and embedding similarity are reported in Table 6, and their per-domain GPT-Neo detection results appear in Table 1.

		Detection (GPT-Neo-2.7B)			Cost and quality		
Domain	n	A_0	A_1	$\Delta A \uparrow$	$\overline{\tau} \downarrow$	Sim.↑	$\eta \uparrow$
CMV	2,040	0.613	0.319	0.293	0.127	0.955	2.319
CS/AI	837	0.980	0.775	0.205	0.091	0.982	2.254
ELI5	2,691	0.610	0.332	0.278	0.122	0.962	2.275
Finance	4,415	0.992	0.969	0.023	0.067	0.986	0.339
HellaSwag	2,189	0.670	0.381	0.290	0.135	0.957	2.140
Medicine	1,301	0.999	0.976	0.023	0.106	0.980	0.214
Open QA	3,378	0.936	0.729	0.207	0.082	0.985	2.536
Reddit ELI5	16,610	0.996	0.955	0.041	0.104	0.977	0.394
ROCStories	2,625	0.625	0.183	0.442	0.185	0.946	2.383
SCI	1,796	0.473	0.233	0.240	0.100	0.972	2.415
SQuAD	1,734	0.468	0.250	0.219	0.090	0.976	2.421
TLDR	2,112	0.636	0.357	0.279	0.111	0.974	2.510
WritingPrompts	2,225	0.594	0.316	0.279	0.158	0.941	1.765
XSum	2,884	0.472	0.247	0.225	0.099	0.973	2.277
Yelp	2,206	0.530	0.212	0.318	0.161	0.948	1.978
Macro mean	–	0.706	0.482	**0.224**	**0.116**	0.968	**1.881**

Table 3: Full-data domain results for WaveHumanize with High-Energy Exclusion. A_0 and A_1 are original and rewritten GPT-Neo-2.7B likelihood AUROC and $\Delta A = A_0 - A_1$. $\overline{\tau}$ is token-sequence dissimilarity, Sim. is embedding similarity, and $\eta = \Delta A / \overline{\tau}$ is editing efficiency. Each domain uses its own human reference pool. The final row is an unweighted mean across domains, not a pooled AUROC. Macro entries are unweighted means over the fifteen domain values.

DIPPER. We use DIPPER-XXL with its official legacy Punkt segmentation, top-$p = 0.75$, and three-sentence windows. The recorded lexical/ordering configuration is `lex_40_order_100`. Long prefixes are left-truncated to 380 tokens. The 30%-budget and full-document settings are separate runs.

G Complete Domain Results

	Likelihood family			Fluctuation family		Other readouts	
Domain	LL	LR	LRR	Spectral	FD	Top-10	Entropy
CMV	+0.293	+0.300	+0.264	+0.142	+0.287	+0.253	+0.188
CS/AI	+0.205	+0.156	+0.080	+0.394	+0.075	+0.106	+0.221
ELI5	+0.278	+0.274	+0.225	+0.132	+0.275	+0.234	+0.163
Finance	+0.023	+0.020	+0.023	+0.181	+0.049	+0.014	+0.035
HellaSwag	+0.290	+0.276	+0.172	+0.094	+0.317	+0.246	+0.167
Medicine	+0.023	+0.017	+0.015	+0.075	+0.180	+0.013	+0.020
Open QA	+0.207	+0.189	+0.131	+0.092	+0.152	+0.162	+0.122
Reddit ELI5	+0.041	+0.034	+0.032	+0.165	+0.162	+0.026	+0.044
ROCStories	+0.442	+0.442	+0.362	+0.197	+0.434	+0.387	+0.276
SCI	+0.240	+0.239	+0.206	+0.156	+0.247	+0.207	+0.152
SQuAD	+0.219	+0.219	+0.168	+0.108	+0.220	+0.186	+0.133
TLDR	+0.279	+0.272	+0.205	+0.121	+0.263	+0.227	+0.177
WritingPrompts	+0.279	+0.273	+0.226	+0.177	+0.273	+0.255	+0.213
XSum	+0.225	+0.216	+0.174	+0.150	+0.219	+0.197	+0.162
Yelp	+0.318	+0.318	+0.269	+0.129	+0.337	+0.278	+0.180
Macro mean	+0.224	+0.216	+0.171	+0.154	+0.228	+0.188	+0.139

Table 4: All seven GPT-Neo criteria for WaveHumanize: per-domain ΔAUROC. LL: likelihood; LR: negative log-rank; FD: analytic self-reference Fast-DetectGPT. Entropy is negated before evaluation. All 105 domain–criterion changes are positive.

	Ours (local)	Full-document baselines		
Domain	ΔA	TP full	AP full	DIPPER full
CMV	0.293	0.361	0.122	-0.181
CS/AI	0.205	0.655	0.604	0.123
ELI5	0.278	0.299	0.118	-0.162
Finance	0.023	0.253	0.197	0.016
HellaSwag	0.290	0.295	0.063	-0.128
Medicine	0.023	0.183	0.078	0.004
Open QA	0.207	0.470	0.447	0.150
Reddit ELI5	0.041	0.225	0.173	0.012
ROCStories	0.442	0.481	0.325	-0.057
SCI	0.240	0.292	0.202	-0.072
SQuAD	0.219	0.287	0.176	-0.074
TLDR	0.279	0.447	0.258	-0.009
WritingPrompts	0.279	0.333	0.143	-0.170
XSum	0.225	0.378	0.264	-0.029
Yelp	0.318	0.285	-0.071	-0.254

Table 5: Full-document baselines as a separate intervention-size comparison: likelihood ΔAUROC by domain. These systems may change substantially more text than the local method; corresponding costs and similarities are in Table 8.

These tables provide the complete per-domain measurements underlying the main-text summaries. Local-budget and full-document baselines are reported separately because their intervention sizes differ substantially. The local-budget comparison addresses the primary setting of reducing detectability with limited editing, whereas the full-document results contextualize the effect of allowing substantially larger interventions.

H Additional Quality and Cost Comparisons

	WaveHumanize (ours)		Simple		Recursive	
Domain	$\overline{\tau}$	Sim.	$\overline{\tau}$	Sim.	$\overline{\tau}$	Sim.
CMV	0.127	0.955	0.292	0.908	0.328	0.882
CS/AI	0.091	0.982	0.254	0.960	0.278	0.945
ELI5	0.122	0.962	0.278	0.923	0.316	0.901
Finance	0.067	0.986	0.235	0.960	0.257	0.947
HellaSwag	0.135	0.957	0.279	0.918	0.325	0.888
Medicine	0.106	0.980	0.255	0.965	0.277	0.950
Open QA	0.082	0.985	0.222	0.967	0.255	0.954
Reddit ELI5	0.104	0.977	0.270	0.945	0.300	0.927
ROCStories	0.185	0.946	0.275	0.923	0.324	0.891
SCI	0.100	0.972	0.246	0.954	0.283	0.937
SQuAD	0.090	0.976	0.231	0.955	0.270	0.938
TLDR	0.111	0.974	0.259	0.951	0.298	0.932
WritingPrompts	0.158	0.941	0.286	0.898	0.327	0.864
XSum	0.099	0.973	0.250	0.947	0.286	0.930
Yelp	0.161	0.948	0.283	0.917	0.334	0.882

Table 6: Newly completed single-pass and two-round local paraphrase comparisons, with all 49,043 machine texts and unchanged outputs retained. Each method uses a matched nominal selection setting. WaveHumanize has lower dissimilarity and higher embedding similarity in every domain. These columns measure cost and quality, not detector scores.

	Local configurations ($\overline{\tau}$/Sim.)				
Domain	TP (N=1)	TP (N=7)	AP	AP (RADAR)	DIPPER
CMV	0.201/0.957	0.152/0.966	0.305/0.898	0.292/0.901	0.153/0.958
CS/AI	0.200/0.978	0.211/0.974	0.285/0.949	0.275/0.956	0.125/0.986
ELI5	0.192/0.960	0.138/0.972	0.302/0.909	0.288/0.906	0.157/0.956
Finance	0.207/0.971	0.205/0.970	0.263/0.951	0.218/0.953	0.114/0.978
HellaSwag	0.187/0.962	0.129/0.974	0.298/0.904	0.292/0.891	0.158/0.949
Medicine	0.218/0.978	0.231/0.975	0.272/0.957	0.264/0.957	0.129/0.980
Open QA	0.166/0.978	0.174/0.977	0.255/0.956	0.238/0.961	0.108/0.979
Reddit ELI5	0.207/0.970	0.223/0.967	0.293/0.936	0.286/0.940	0.122/0.979
ROCStories	0.166/0.970	0.133/0.974	0.311/0.902	0.292/0.904	0.133/0.966
SCI	0.178/0.975	0.121/0.982	0.281/0.936	0.262/0.941	0.150/0.967
SQuAD	0.168/0.974	0.123/0.980	0.254/0.945	0.235/0.948	0.147/0.967
TLDR	0.194/0.969	0.170/0.971	0.280/0.939	0.271/0.936	0.154/0.967
WritingPrompts	0.180/0.959	0.123/0.973	0.314/0.877	0.312/0.876	0.145/0.955
XSum	0.186/0.966	0.154/0.971	0.277/0.932	0.252/0.941	0.156/0.956
Yelp	0.179/0.965	0.113/0.979	0.305/0.901	0.298/0.899	0.150/0.954

Table 7: Additional local configurations: each cell is $\overline{\tau}$/embedding similarity. All columns cover the full per-domain machine pools.

	Full configurations ($\overline{\tau}$/Sim.)		
Domain	TP	AP	DIPPER
CMV	0.625/0.870	0.809/0.742	0.488/0.863
CS/AI	0.639/0.925	0.733/0.845	0.445/0.929
ELI5	0.585/0.881	0.790/0.757	0.513/0.853
Finance	0.693/0.906	0.719/0.844	0.490/0.891
HellaSwag	0.573/0.892	0.778/0.756	0.529/0.831
Medicine	0.687/0.928	0.693/0.868	0.467/0.910
Open QA	0.541/0.929	0.682/0.845	0.440/0.898
Reddit ELI5	0.685/0.903	0.785/0.801	0.486/0.898
ROCStories	0.560/0.910	0.824/0.749	0.504/0.859
SCI	0.536/0.930	0.697/0.857	0.497/0.883
SQuAD	0.536/0.917	0.696/0.847	0.495/0.886
TLDR	0.598/0.912	0.730/0.850	0.516/0.885
WritingPrompts	0.603/0.875	0.839/0.677	0.510/0.843
XSum	0.566/0.909	0.733/0.818	0.534/0.852
Yelp	0.562/0.895	0.812/0.741	0.508/0.844

Table 8: Additional full configurations: each cell is $\overline{\tau}$/embedding similarity, covering all 49,043 machine documents.

Embedding similarity is computed with `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019). It is used solely as an evaluation measure of semantic proximity between source and rewritten documents; WaveHumanize does not use an embedding-similarity gate during localization, candidate generation, or candidate selection.

Domain	GPT-Neo-2.7B	llm-content-detector	chatgpt-detector
CMV	0.293	0.230	0.082
CS/AI	0.205	0.101	0.017
ELI5	0.278	0.209	0.113
Finance	0.023	0.005	0.001
HellaSwag	0.289	0.280	0.119
Medicine	0.023	0.005	0.001
Open QA	0.207	0.128	0.016
Reddit ELI5	0.041	0.002	0.000
ROCStories	**0.442**	**0.422**	**0.300**
SCI	0.240	0.246	0.102
SQuAD	0.219	0.244	0.110
TLDR	0.279	0.341	0.222
WritingPrompts	0.279	0.231	0.176
XSum	0.225	0.291	0.168
Yelp	0.318	0.274	0.139
Macro mean	**0.224**	**0.201**	**0.104**

Table 10: Per-domain detectability reduction $\Delta A = A_0 - A_1$ for High-Energy Exclusion (Ours) under three evaluators that do not use the GPT-J localization proxy, all rescored offline on frozen outputs. Every entry is non-negative. The final row is an unweighted mean over the fifteen domains.

I Transfer across Independent Detector Families

Detector	Family	MAGE			HC3		
		A_0	A_1	$\Delta A \uparrow$	A_0	A_1	$\Delta A \uparrow$
Likelihood family							
GPT-Neo-2.7B (likelihood)	LM-family likelihood	0.555	0.275	0.280	0.994	0.951	0.043
Supervised family							
chatgpt-detector	Supervised (RoBERTa)	0.569	0.435	0.134	1.000	0.999	0.001
llm-content-detector	Supervised (RoBERTa)	0.651	0.399	0.253	1.000	0.994	0.005
openai_base	Supervised (RoBERTa-base)	0.733	0.670	0.064	0.990	0.671	**0.319**
openai_large	Supervised (RoBERTa-large)	0.770	0.748	0.022	0.974	0.683	0.292
desklib	Supervised (DebertaV3-large)	0.917	0.842	0.075	0.999	0.977	0.022
Adversarial and reference-free							
RADAR	Generative adversarial	0.665	0.700	−0.036	0.606	0.599	0.007
binoculars	Reference-free (Llama-3-8B)	0.559	0.364	0.195	0.993	0.865	0.128

Table 9: Pooled per-corpus transfer of WaveHumanize with High-Energy Exclusion, rescored offline on frozen outputs (MAGE $n = 22{,}502$; HC3 $n = 26{,}541$) by detectors independent of the GPT-J proxy. A_0/A_1 are original/rewritten AUROC and $\Delta A = A_0 - A_1$. Bold marks the largest positive transfer. Supervised and reference-free detectors truncate at 512 tokens (likelihood columns at 2048); affected proportions are below 3%. chatgpt-detector and llm-content-detector are trained on HC3-derived data, so their HC3 rows have near-ceiling original AUROC (1.000). RADAR scores exclude paraphrase-guided baselines (oracle conflict). Interpretation and the desklib boundary are given in Section 4.4.

All transfer measurements in Table 9 are obtained by rescoring the frozen main-configuration outputs offline; no transfer detector is queried during rewriting. Results are pooled separately within MAGE and HC3 so that different detector families are evaluated on identical text pools. This analysis is distinct from the seven GPT-Neo scoring criteria, which share a single evaluator checkpoint.

J Independent-Detector Matrices

Table 10 reports per-domain reductions under the independent GPT-Neo-2.7B likelihood evaluator and two supervised detectors. All reported entries are non-negative.

Table 11 compares all 30%-budget configurations under the same detector set, holding the rewriter, fragment width, candidate count, round limit, and nominal budget fixed; High-Energy Exclusion matches or exceeds Random localization across these detector–corpus pairs. High-Energy Exclusion achieves the largest reductions on the likelihood-family and supervised detectors in both corpora,

MAGE								
Configuration	GPT-Neo	chatgpt	llm-cont	RADAR	oBase	oLarge	desklib	bino
Locally guided								
High-Energy Exclusion (Ours)	**0.280**	**0.134**	**0.253**	-0.036	0.064	0.022	**0.075**	**0.195**
Random	0.270	0.130	0.241	-0.041	0.058	0.019	0.065	0.175
Unguided paraphrase								
Simple	-0.142	-0.163	-0.222	-0.032	-0.001	0.003	-0.069	-0.352
Recursive	-0.093	-0.130	-0.206	-0.026	0.053	0.039	-0.073	-0.350
TP (N=1)	0.110	0.047	0.123	-0.006	-0.008	-0.036	0.018	0.064
TP (N=7)	0.075	0.090	0.110	**0.003**	0.000	-0.028	0.008	0.041
DIPPER	-0.023	-0.022	-0.011	-0.007	-0.027	-0.030	0.015	-0.038
Detector-guided paraphrase								
AP	0.141	0.116	0.107	—	0.157	0.091	0.005	0.015
AP (RADAR)	0.135	0.120	0.094	—	**0.159**	**0.093**	0.002	0.016

HC3								
Configuration	GPT-Neo	chatgpt	llm-cont	RADAR	oBase	oLarge	desklib	bino
Locally guided								
High-Energy Exclusion (Ours)	**0.043**	**0.001**	**0.005**	0.007	**0.319**	**0.292**	**0.022**	**0.128**
Random	0.040	**0.001**	**0.005**	-0.005	0.296	0.279	0.019	0.114
Unguided paraphrase								
Simple	0.021	0.000	0.001	0.027	0.140	0.159	-0.001	-0.002
Recursive	0.034	0.000	0.001	0.030	0.174	0.177	-0.001	-0.001
TP (N=1)	0.014	0.000	0.002	0.073	0.053	0.067	0.003	0.032
TP (N=7)	0.014	0.000	0.002	**0.093**	0.042	0.044	0.002	0.030
DIPPER	0.002	0.000	0.000	0.025	0.005	0.008	0.001	0.010
Detector-guided paraphrase								
AP	0.025	**0.001**	0.004	—	0.270	0.262	0.002	0.040
AP (RADAR)	0.024	**0.001**	0.003	—	0.261	0.254	0.001	0.043

Table 11: Same-budget comparison across independent detectors: pooled $\Delta A = A_0 - A_1$ for every budget-matched local configuration under each detector's primary criterion. Bold marks the best value in a column; where several configurations tie at the maximum, all of them are bolded. Detector abbreviations: GPT-Neo-2.7B, chatgpt-detector, llm-content-detector, RADAR, openai-base, openai-large, desklib, and binoculars. All configurations operate under a matched local-edit budget; realized token change differs and is reported in the main text. AP rows are excluded under RADAR, which is AP's oracle. All numbers are offline rescorings of frozen outputs; supervised and reference-free detectors truncate at 512 tokens.

Representation	AUROC	AUPRC	$\text{AUROC}_{\text{int}}$	$\text{AUPRC}_{\text{int}}$
STFT	0.739	0.742	0.742	0.734
Haar approx. (ours)	**0.774**	**0.758**	**0.777**	**0.751**

Table 12: Summary of the controlled span-localization diagnostic, macro-averaged over the four corpus–proxy settings. Interior metrics exclude tokens adjacent to the human–machine transition. This is the compact form of Table 13.

with a single exception: on MAGE the two OpenAI criteria are led by Adversarial Paraphrasing (0.157 and 0.159, against 0.064 and 0.022 for WaveHumanize).

That baseline is also the strongest comparator overall, and the reason is structural rather than incidental. It is the only compared method that receives detector feedback while rewriting, because its decoding is steered by RADAR (Cheng et al., 2025; Hu et al., 2023); it can therefore search directly for paraphrases that lower a learned detector score, whereas WaveHumanize is restricted to the proxy statistics of one fixed language model. Its advantage is nonetheless bounded: Adversarial Paraphrasing loses to WaveHumanize on every other criterion in Table 11, on the likelihood-family detectors of both corpora, and throughout HC3, and it inherits the cost of requiring a detector oracle, which the setting of this work excludes. Full-document configurations are reported separately in Appendix G.

K Localization and Candidate-Ranking Analysis

This section expands the three component analyses summarized in Section 4: the local representation, the localization policy, and candidate ranking. We then examine robustness to alternative estimators of local statistical fluctuation.

Representation diagnostic. We first isolate the representation itself, without invoking the rewriting model or candidate selection. Each diagnostic document contains a human-written prefix followed by a machine-generated continuation whose boundary is known. A local representation is evaluated by how well its token scores rank the machine-generated span above the human prefix.

We report token-level AUROC and AUPRC, together with interior variants that remove tokens adjacent to the human–machine transition. This separates within-span localization ability from artifacts caused by the splice boundary.

Across both corpora and both proxy models, the Haar approximation-energy profile consistently localizes the machine-generated span more accurately than matched-window STFT. Averaged over the four corpus–proxy settings, token AUROC increases from 0.739 to 0.774 and AUPRC from 0.742 to 0.758. After boundary tokens are removed, interior AUROC increases from 0.742 to 0.777 and interior AUPRC from 0.734 to 0.751. The similarity of the full-span and interior gaps indicates that the advantage is not primarily driven by splice-boundary artifacts.

Paired per-document bootstrap tests give Haar–STFT AUPRC differences of $+0.012$ to $+0.024$, with all 95% intervals above zero. Under the deployed 30% high-energy exclusion rule, Haar retains 88.7% of machine-region tokens while excluding 60.3% of human-region tokens, compared with 86.2% and 56.0% respectively for STFT.

Full-data localization-policy comparison. High-Energy Exclusion, Lowest-Energy, and Random share the same rewriter, fragment width, candidate count, round limit, and nominal selection budget; only the localization policy changes. Table 14 evaluates their frozen outputs with the independent GPT-Neo likelihood score and reports both paired score differences and realized edit-cost differences.

High-Energy Exclusion yields larger score reductions than Lowest-Energy in all fifteen domains, with every paired 95% confidence interval above zero. The score advantage ranges from 0.024 to 0.052, while the corresponding change in domain-level mean token-sequence dissimilarity is at most 0.0051. Relative to Random localization, all point estimates favor High-Energy Exclusion and the realized edit costs remain closely matched.

	MAGE				HC3			
	AUROC		AUPRC		AUROC		AUPRC	
Representation	GPT-J	GPT-Neo	GPT-J	GPT-Neo	GPT-J	GPT-Neo	GPT-J	GPT-Neo
Random (sanity)	0.500	0.499	0.583	0.582	0.503	0.503	0.582	0.581
STFT	0.741	0.731	0.744	0.737	0.745	0.741	0.745	0.743
Haar approximation (ours)	**0.773**	**0.764**	**0.757**	**0.749**	**0.785**	**0.774**	**0.769**	**0.758**

Table 13: Localization ability of local representations. Each document is a human prompt continued by an instruction-tuned model; the task is to rank the tokens of the continued machine-written span above the human prefix, evaluated per document by token AUROC and AUPRC (macro averages). All representations use a matched effective span of about ten tokens. The two proxies are reported separately. Every local representation is far above the random-profile sanity check, and the Haar approximation exceeds a matched-window STFT. The random AUPRC reflects the machine-token prevalence in the constructed sequences. Bold marks the best representation per column.

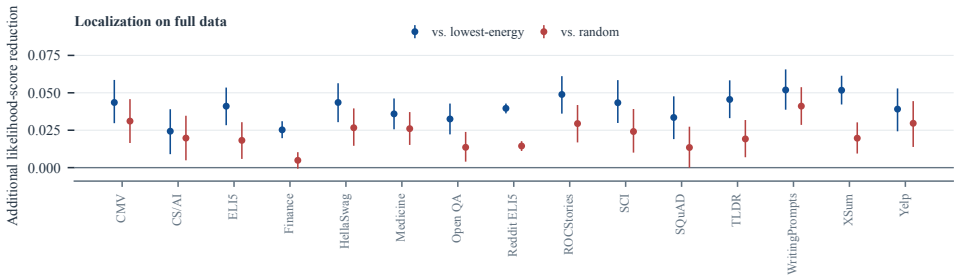

Figure 6: Complementary evidence for localization. Paired GPT-Neo likelihood-score gains of High-Energy Exclusion over Lowest-Energy and Random localization on the full data, with 95% document-bootstrap intervals.

Deployed candidate-ranking ablation. We next isolate the ranking rule by fixing the High-Energy Exclusion locator, rewriter, candidate pool, and round limit, and replacing only the deployed Haar approximation-energy ranking with uniform random selection over the same candidates.

Energy-based ranking reduces independent GPT-Neo-2.7B likelihood AUROC by 0.277 on MAGE and 0.116 on HC3, compared with 0.220 and 0.078 under random candidate selection. The energy-ranked arm is stronger in all fifteen domains, corresponding to corpus-level gaps of 0.057 and 0.037. Realized edit cost remains closely matched: across domains, token-sequence dissimilarity differs by at most 0.009 and embedding similarity by at most 0.008.

Localization-rule attribution. The diagnostic attribution experiment varies the localization rule while holding the candidate-ranking rule fixed. Highest-Energy localization selects windows far outside the energy regime retained by High-Energy Exclusion and yields the weakest escape results in both corpora. These diagnostic-subset measurements are reported separately from the full-data localization results.

Robustness to the local fluctuation estimator. The localized-editing effect does not depend strongly on a single wavelet band. Level-one approximation energy, level-one and level-two detail energies, and their multi-scale fusion produce similar reductions on the domain-balanced diagnostic subset.

The same conclusion extends beyond wavelet coefficients: matched raw-variance and sliding-window-variance estimators also remain well above Random and produce reductions in the same general range. In particular, matching the effective analysis span makes the wavelet-detail and raw-variance paths numerically identical in the recorded diagnostic runs. These results suggest that the broader phenomenon is the concentration of local statistical fluctuation, while the deployed level-one Haar approximation provides a compact, position-resolved, and interpretable estimator of that structure.

Domain	High-Energy Exclusion (Ours) − Lowest-Energy			High-Energy Exclusion (Ours) − Random		
	Score diff.↑	95% CI	$\Delta\tau$	Score diff.↑	95% CI	$\Delta\tau$
CMV	+0.044	[+0.030, +0.059]	-0.00037	+0.031	[+0.016, +0.046]	+0.005
CS/AI	+0.024	[+0.009, +0.039]	-0.002	+0.020	[+0.005, +0.035]	+0.003
ELI5	+0.041	[+0.028, +0.053]	-0.001	+0.018	[+0.006, +0.030]	+0.003
Finance	+0.025	[+0.020, +0.031]	+0.00017	+0.005	[-0.001, +0.010]	+0.001
HellaSwag	+0.044	[+0.030, +0.056]	-0.003	+0.027	[+0.015, +0.040]	+0.006
Medicine	+0.036	[+0.026, +0.046]	-0.001	+0.026	[+0.015, +0.037]	+0.005
Open QA	+0.033	[+0.022, +0.043]	-0.001	+0.014	[+0.004, +0.024]	+0.003
Reddit ELI5	+0.040	[+0.036, +0.043]	+0.001	+0.014	[+0.011, +0.018]	+0.004
ROCStories	+0.049	[+0.036, +0.061]	-0.005	+0.029	[+0.017, +0.042]	+0.006
SCI	+0.043	[+0.030, +0.058]	-0.001	+0.024	[+0.010, +0.039]	+0.003
SQuAD	+0.034	[+0.019, +0.048]	-0.002	+0.013	[-0.00003, +0.027]	+0.003
TLDR	+0.046	[+0.033, +0.058]	+0.002	+0.019	[+0.007, +0.032]	+0.004
WritingPrompts	+0.052	[+0.039, +0.066]	+0.00019	+0.041	[+0.029, +0.054]	+0.007
XSum	+0.052	[+0.042, +0.061]	+0.001	+0.020	[+0.009, +0.030]	+0.005
Yelp	+0.039	[+0.024, +0.053]	-0.005	+0.030	[+0.014, +0.044]	+0.007
Macro mean	+0.040		-0.0012	+0.023		+0.004

Table 14: Full-data paired localization contrasts under independent GPT-Neo likelihood, comparing High-Energy Exclusion (Ours) against two alternative placement rules, Lowest-Energy and Random. Each block reports the paired score difference (alternative minus High-Energy Exclusion (Ours); larger favors the proposed rule), its 95% resampled interval, and the associated $\Delta\tau$ (High-Energy Exclusion (Ours) minus alternative; negative favors the proposed rule). Confidence intervals use 2,000 paired document resamples and are not adjusted for multiple domain comparisons. They concern score differences, not AUROC.

Domain	ΔA_{en}	ΔA_{rnd}	Δ	$\overline{\tau}$	Sim.
CMV	0.266	0.200	0.067	0.125	0.954
CS/AI	0.219	0.138	0.081	0.092	0.982
ELI5	0.282	0.220	0.062	0.123	0.961
Finance	0.028	0.018	0.011	0.068	0.984
HellaSwag	0.314	0.239	0.076	0.145	0.953
Medicine	0.022	0.009	0.013	0.106	0.978
Open QA	0.213	0.160	0.053	0.083	0.987
Reddit ELI5	0.042	0.024	0.018	0.101	0.978
ROCStories	0.422	0.345	0.077	0.179	0.947
SCI	0.250	0.197	0.053	0.097	0.973
SQuAD	0.201	0.175	0.027	0.093	0.977
TLDR	0.278	0.217	0.061	0.116	0.975
WritingPrompts	0.308	0.242	0.066	0.156	0.943
XSum	0.217	0.157	0.061	0.099	0.975
Yelp	0.329	0.265	0.064	0.167	0.950
Macro mean	0.226	0.174	0.052	0.117	0.968

Table 15: Candidate-ranking ablation on the domain-balanced 2000-pair subset. Holding the High-Energy Exclusion locator, the generator, the candidate pool, and the round limit fixed, energy-based candidate ranking is compared with uniform random selection over the same candidates, so only the ranking rule changes. ΔA_{en} and ΔA_{rnd} are reductions in independent GPT-Neo-2.7B likelihood AUROC (Section 4.1); Δ is their difference, $\overline{\tau}$ is token-sequence dissimilarity and Sim. embedding similarity, both from the energy-ranked arm. Random selection matches $\overline{\tau}$ to within 0.009 and Sim. to within 0.008 in every domain. The final row is the unweighted macro mean over the fifteen domains.

L Per-Domain Operating Points

All fixed thresholds are calibrated independently within each domain and then held unchanged for original and rewritten machine texts. This evaluation complements AUROC by measuring detector decisions under a fixed operating rule. The evaluator is never queried during rewriting.

Opposite-tail control. Because AUROC is directional, an AUROC below 0.5 could in principle arise if rewriting pushes machine-generated texts past the human distribution and into the opposite score tail. We therefore retain the reported upper-tail threshold and additionally flag texts below the human 0.5% calibration quantile, giving a nominal total calibration FPR of approximately 1.5%.

	MAGE		HC3	
Localization rule	sel. energy	escape	sel. energy	escape
Energy-guided localization				
Lowest-Energy	1.02	0.258	0.29	0.128
Random	3.31	0.264	1.53	0.137
Variance	2.24	0.270	0.77	0.140
High-energy control				
Highest-Energy	13.02	0.182	13.61	0.083

Table 16: Localization-rule attribution on the domain-balanced 2000-pair diagnostic subset (fulltext-window pipeline, GPT-J-6B proxy, likelihood escape = AUROC drop). All arms use the deployed energy-max candidate ranking and vary only the localization rule. "sel. energy" is the mean local energy of the selected windows, a manipulation check; escape is the primary endpoint. The Highest-Energy control selects windows far outside the energy range of High-Energy Exclusion (Ours) and yields the lowest escape. These subset values are diagnostic and are not pooled with the full-data results.

Estimator family	Energy estimator	MAGE $\Delta A \uparrow$	HC3 $\Delta A \uparrow$
Wavelet profile			
	Haar level-2 approximation (cA_2)	0.244	0.127
	Haar level-1 detail (cD_1)	0.262	0.129
	Haar level-2 detail (cD_2)	0.256	0.127
	multi-scale fusion (MS_2)	0.261	0.130
Reference			
	random selection	<u>0.210</u>	<u>0.094</u>

Table 17: Which wavelet band supplies the localized-editing signal, on the domain-balanced 2,000-pair diagnostic subset used for the attribution analysis (fulltext-window pipeline, GPT-J-6B proxy, likelihood AUROC). Approximate reduction is stable across the level-one approximation, the level-one and level-two detail bands, and their multi-scale fusion, and every configuration remains far above random window selection. The localized-editing benefit therefore follows from editing where local fluctuation concentrates; the wavelet profile provides an interpretable, parameter-free instantiation of that target.

Across all fifteen domains, $15{,}651$ rewritten texts escape the original upper-tail criterion. The additional lower-tail rule recaptures only 11 of these texts, or 0.07%. A symmetric two-sided threshold yields the same qualitative conclusion. Thus, the reduction observed under the fixed upper-tail operating point is not primarily explained by rewrites crossing into the opposite extreme of the score distribution.

The two ASR columns in the full table use different original-detection sets because the two-sided rule also flags some original texts through the lower tail; they should therefore not be interpreted as directly comparable denominators.

M ALGORITHM

Fixed-threshold validation	Result
Macro conditional ASR at 1% FPR	80.8%
Best domain (CS/AI)	**98.9%**
One-sided escapes	15,651
Recaptured by the lower tail	11 (0.07%)

Table 18: Summary of the fixed-threshold operating point under the independent GPT-Neo-2.7B likelihood evaluator. Per-domain values are given in Table 19 and Table 20.

		Held-out decisions		TPR		Conditional success	
Domain	n_H	$n_{H,\mathrm{test}}$	FPR	orig.	rew.	escaped/detected	ASR↑
CMV	1,907	953	0.007	0.315	0.033	575/642	0.896
CS/AI	764	382	0.005	0.337	0.004	279/282	0.989
ELI5	2,415	1,207	0.009	0.212	0.022	510/570	0.895
Finance	3,392	1,696	0.009	0.980	0.726	1122/4328	0.259
HellaSwag	1,337	669	0.004	0.197	0.014	400/431	0.928
Medicine	1,171	585	0.029	0.998	0.778	286/1298	0.220
Open QA	515	257	0.004	0.128	0.015	382/433	0.882
Reddit ELI5	40,326	20,163	0.011	0.983	0.432	9150/16320	0.561
ROCStories	3,109	1,555	0.009	0.240	0.004	620/630	0.984
SCI	2,523	1,261	0.015	0.173	0.018	278/311	0.894
SQuAD	2,484	1,242	0.011	0.066	0.012	95/115	0.826
TLDR	2,429	1,215	0.006	0.257	0.023	495/543	0.912
WritingPrompts	1,414	707	0.016	0.302	0.009	652/671	0.972
XSum	2,150	1,075	0.002	0.185	0.010	505/534	0.946
Yelp	2,250	1,125	0.015	0.143	0.006	302/315	0.959

Table 19: Independent GPT-Neo likelihood at a within-domain threshold targeting 1% calibration FPR. n_H includes calibration and held-out humans; Test FPR is measured only on the held-out portion. Conditional ASR counts only machine texts detected before rewriting. No detector threshold is retuned after rewriting.

Domain	n_1	One-sided ASR	Escapes	n_2	Two-sided ASR	Recaptured
CMV	642	0.896	575	723	0.795	0.35%
CS/AI	282	0.989	279	282	0.989	0.00%
ELI5	570	0.895	510	614	0.832	0.00%
Finance	4,328	0.259	1,122	4,329	0.259	0.00%
HellaSwag	431	0.928	400	445	0.899	0.00%
Medicine	1,298	0.220	286	1,298	0.220	0.00%
Open QA	433	0.882	382	433	0.882	0.00%
Reddit ELI5	16,320	0.561	9,150	16,320	0.561	0.00%
ROCStories	630	0.984	620	694	0.883	1.13%
SCI	311	0.894	278	427	0.651	0.36%
SQuAD	115	0.826	95	221	0.430	0.00%
TLDR	543	0.912	495	568	0.870	0.20%
WritingPrompts	671	0.972	652	708	0.924	0.00%
XSum	534	0.946	505	743	0.681	0.00%
Yelp	315	0.959	302	358	0.849	0.00%
All domains	27,423		15,651	28,163		0.07%

Table 20: Per-domain two-sided fixed-threshold analysis under the main scorer (GPT-Neo-2.7B likelihood). n_1 is the number of original texts detected by the 1% upper threshold of Section 4.5, and "One-sided ASR" is the fraction of those that are no longer detected after rewriting ("Escapes"). n_2 applies the same upper threshold together with an additional lower-tail criterion at the human 0.5% quantile; because that criterion also flags additional original texts, $n_2 \geq n_1$ and the two ASR columns are *not* like-for-like. "Recaptured" is the like-for-like quantity: the fraction of the n_1 escapes that the added lower-tail criterion catches. The final row sums both denominators and the escapes.

Algorithm 1 WaveHumanize with High-Energy Exclusion

Require: Document x, proxy model p_θ, rewriter q_ϕ; α, ρ, W, K, R

for $r = 1, \ldots, R$ **do**
  Compute the local energy profile $E(x)$
  Construct candidate windows $\mathcal{W}(x)$ and exclude the high-energy tail to obtain $\mathcal{P}(x)$
  Sample non-overlapping edit spans $\mathcal{S}$ from $\mathcal{P}(x)$ within budget B
  $\mathcal{U} \leftarrow \varnothing$
  for each $s \in \mathcal{S}$ **do**
    Generate and filter K contextual rewrites $\mathcal{C}_s$ with q_ϕ
    $c^* \leftarrow \arg\max_{c \in \mathcal{C}_s} G(c; x, s)$
    Add (s, c^*) to $\mathcal{U}$ if $\mathcal{C}_s \neq \varnothing$
  end for
  if $\mathcal{U} = \varnothing$ **then**
    break
  end if
  Apply $\mathcal{U}$ to x from right to left
end for
return x ▷ rewritten document