ATTACKING AI EDITING SCORES WITH ADVERSARIAL REWRITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) now generate and revise text so rapidly that verifying its provenance has become increasingly difficult across scholarship, journalism, law, and creative writing. The written record is becoming a blend of human and machine contributions, and the central question is no longer whether AI was involved, but how much of a text it shaped and how reliably that can be measured. Models designed for this task assign AI editing scores, and because such scores increasingly inform judgments about authorship, credit, and integrity, their trustworthiness under adversarial pressure is a critical concern. In practice, however, the original human-written text is usually unavailable when a text is scored, which may make AI editing scores sensitive to surface form rather than the actual extent of AI editing. We study whether meaning-preserving rewrites can substantially reduce AI editing scores in this setting. We propose Prior-guided Rewriting for Editing Score Suppression (PRESS), which uses a paired humanization prior to identify regions characteristic of AI-edited writing and a surrogate model to select candidate rewrites. PRESS requires neither target model queries nor the original human-written text paired with the input, while a perturbation budget and quality checks control the rewriting process. The experimental results show that PRESS can lower AI editing scores across four models without moving the adversarial text closer to the original human-written text, raising concerns about their reliability as measures of the extent of AI editing.

1 INTRODUCTION

Large language models (LLMs) are increasingly embedded across scientific communication, journalism, education, and professional writing, and they no longer merely generate complete texts: they participate throughout the writing process, from grammar correction and polishing to sentence rewriting and more extensive revision. In such cases, simply determining whether AI was involved does not fully describe its role in the writing process. We also need to know how much the text has been changed by AI. Recent work moves beyond binary AI-generated text detection and measures the extent of AI editing. A small correction and a major rewrite both involve AI, but they reflect very different amounts of AI editing. Polish Ratio measures how much ChatGPT changes the original human-written text (Yang et al., 2024). EditLens uses similarity between the original human-written text and AI-edited text as supervision, and trains a regression model that takes only the AI-edited text as input and predicts the amount of AI editing (Thai et al., 2026).

Measuring the extent of AI editing is important, but as these scores increasingly inform judgments about authorship, credit, and integrity, their reliability becomes a practical necessity rather than a technical detail. A lower AI editing score is meant to indicate less AI editing, rather than simply a different way of expressing the same content. For an AI-edited text, a genuine reduction may come from reversing some changes introduced by AI and moving the text closer to its original human-written form. However, the original human-written text is usually unavailable when a new text is evaluated in practice. The model therefore estimates the extent of AI editing from the evaluated text alone, which may make the score sensitive to changes in surface form. **This setting motivates a robustness study of whether meaning-preserving rewrites can substantially reduce AI editing scores without access to the original human-written text**.

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.

Similar attacks have been widely studied for AI detection. Prior work shows that paraphrasing can weaken AI detection, while other methods use word substitution or detector-guided generation to change detection results (Krishna et al., 2023; Sadasivan et al., 2025; Shi et al., 2024; Nicks et al., 2024). More recent methods learn human-like rewriting patterns or design attacks that transfer across different detectors (Xu et al., 2026; Ranganath & Ramesh, 2026). These attacks mainly target detection results or detection confidence, rather than a continuous AI editing score. The robustness of such scores under rewriting remains less understood. To address this gap, we propose Prior-guided Rewriting for Editing Score Suppression (PRESS). PRESS neither queries the target models during attack generation nor uses the original human-written text paired with the current input. Instead, it constructs a paired humanization prior that captures recurring distributional differences between original human-written texts and their corresponding AI-edited texts. The prior helps PRESS identify which parts of the text to rewrite and which candidate rewrites better follow human writing patterns. A surrogate model then helps select candidates that are more likely to reduce the AI editing score. We control the amount of rewriting with a fixed perturbation budget and evaluate the resulting adversarial texts on different target models to test transferability. Based on this framework, our main contributions are:

1. The first attack on AI editing scores. To the best of our knowledge, we are the first to study adversarial rewriting against AI editing scores. We propose PRESS, which reduces AI editing scores without querying the target models or using the original human-written text paired with the current input.

2. Paired humanization prior with finite-sample estimation guarantees. We construct a paired humanization prior to capture distributional differences between original human-written texts and their corresponding AI-edited texts, and estimate it from finite training pairs. Theoretical analysis shows that the estimated prior selects the same candidate as the population prior when the candidates are well separated, and otherwise selects a candidate that remains close to the population-prior choice.

3. Transfer across models and domains. PRESS reduces AI editing scores across four models and also lowers the Fast-DetectGPT score, showing transfer to a different detection objective. Moreover, as shown in Figure 1, the adversarial texts generated by PRESS can receive lower AI editing scores without moving closer to the original human-written text. Such behavior provides evidence that AI editing scores can be unstable under adversarial rewriting.

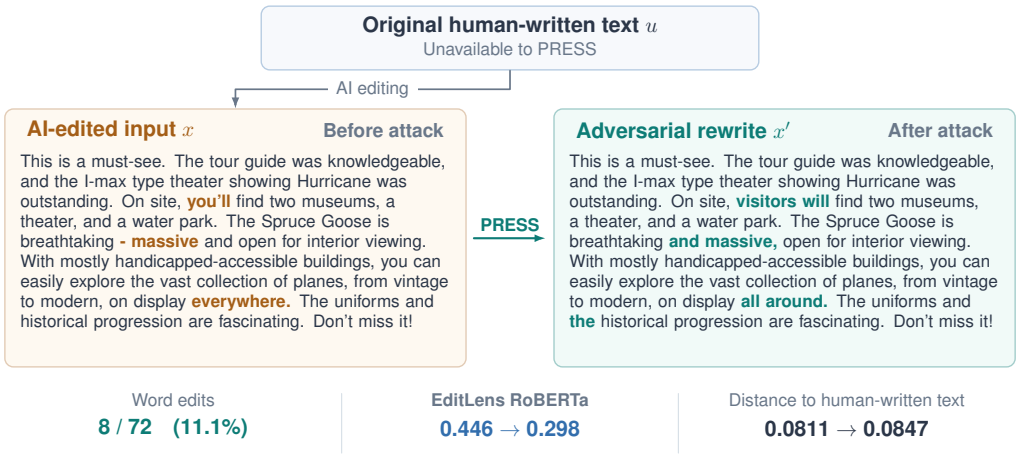


Figure 1: An example of score reduction with PRESS. PRESS rewrites the AI-edited input x into x' without access to the original human-written text u. Highlighted spans mark the changes. PRESS lowers the EditLens RoBERTa score from 0.446 to 0.298, a 33.2% relative reduction. Meanwhile, the distance to u increases from 0.0811 to 0.0847, which shows that a lower AI editing score does not necessarily indicate greater similarity to the original human-written text.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 presents PRESS and the theoretical analysis. Section 4 reports the experimental results. Section 5 concludes the paper.

2 Related Work

Existing work on AI detection uses learned classifiers or statistical signals from language models to identify AI-generated text (Ippolito et al., 2020; Mitchell et al., 2023; Hans et al., 2024). Fast-DetectGPT provides a continuous detection criterion based on language-model probabilities to distinguish machine-generated from human-written text (Bao et al., 2024). Other studies consider mixed authorship, polishing, and different forms of AI-assisted writing (Zhang et al., 2024; Saha & Feizi, 2025; Lei et al., 2025; Li et al., 2026). Polish Ratio and EditLens further quantify the extent of AI editing rather than only making a detection decision (Yang et al., 2024; Thai et al., 2026).

The robustness of AI detection systems has also been studied through rewriting attacks. Early methods such as TextFooler (Jin et al., 2020) and BERT-ATTACK (Li et al., 2020) replace selected words with constrained alternatives. Later, DIPPER (Krishna et al., 2023) uses controlled paraphrasing, while recursive paraphrasing (Sadasivan et al., 2025) repeatedly rewrites the input to make AI-generated text harder to detect. The T5 paraphraser (Vorobev & Kuznetsov, 2023) rewrites each sentence with a paraphrased alternative. More recent methods use detector feedback to guide rewriting or train rewriting models to produce more human-like text (Shi et al., 2024; Nicks et al., 2024; Ranganath & Ramesh, 2026; Xu et al., 2026).

3 Methodology

In this section, we introduce Prior-guided Rewriting for Editing Score Suppression (PRESS), which rewrites the input text to reduce its AI editing score while preserving meaning and controlling the amount of change. First, we define the attack objective and threat model. Then we introduce the paired humanization prior, which guides where the input should be rewritten. After identifying the text regions to edit, PRESS generates candidate rewrites and uses the paired humanization prior together with the surrogate model to rank and select them. Finally, we describe the quality checks and iterative update procedure used to obtain the adversarial text.

3.1 Problem Setup and Threat Model

Let x be the original input, which may be AI-edited or AI-generated, and let $S_T(\cdot)$ denote the target scoring function, with higher scores indicating a greater predicted extent of AI editing. We seek an adversarial text x' that minimizes the target score under a perturbation budget and quality constraints:

$$\min_{x'} S_T(x') \quad \text{subject to} \quad E(x, x') \leq \epsilon, \qquad x' \in \mathcal{C}(x), \tag{1}$$

where $E(x, x')$ is the word-level perturbation rate from x to x', and ϵ denotes the perturbation budget. The set $\mathcal{C}(x)$ contains rewrites that satisfy semantic preservation, content preservation, and grammaticality constraints. During attack generation, PRESS neither queries the target model nor requires the original human-written text u paired with the current input x. Therefore, the target scores, model parameters, and gradients are unavailable during generation.

3.2 Phrase Localization with the Paired Humanization Prior

To satisfy the perturbation budget $E(x, x') \leq \epsilon$, PRESS first determines which parts of the input should be rewritten. Since the original human-written text paired with the current input is unavailable, we cannot directly recover which parts were changed by AI. Instead, we use the paired humanization prior to identify regions whose wording is more characteristic of AI-edited text and prioritize them for rewriting.

Lexical salience Let $\mathcal{P}_{\text{pair}}$ denote the population distribution of paired texts (U, V), where U is an original human-written text and V is the corresponding AI-edited text. The paired humanization prior captures systematic differences between the two sides of $\mathcal{P}_{\text{pair}}$, with lexical salience measuring how characteristic each word is of original human-written or AI-edited text. For a word w, let $p_H(w)$ and $p_A(w)$ denote the population occurrence probabilities on the original human-written

and AI-edited sides of $\mathcal{P}_{\text{pair}}$, respectively. We define the population lexical salience score as

$$\ell_*(w) = \log \frac{p_H(w)}{p_A(w)}. \tag{2}$$

A positive value indicates that w is more characteristic of original human-written text, while a negative value indicates that it is more characteristic of AI-edited text.

Since $\mathcal{P}_{\text{pair}}$ is unknown in practice, we estimate these probabilities from a finite paired sample $\mathcal{D}_{\text{pair}} = \{(u_i, v_i)\}_{i=1}^{N}$, where each u_i is an original human-written text and v_i is the corresponding AI-edited text. Let $C_H(w)$ and $C_A(w)$ be the total numbers of times word w appears in the original human-written texts and AI-edited texts, respectively. Let T_H and T_A be the total numbers of words on the two sides of the paired sample. After removing infrequent words, let $\mathcal{V}$ denote the remaining vocabulary. We then estimate $p_H(w)$ and $p_A(w)$ as

$$\widehat{p}_H(w) = \frac{C_H(w) + a}{T_H + a|\mathcal{V}|}, \qquad \widehat{p}_A(w) = \frac{C_A(w) + a}{T_A + a|\mathcal{V}|}. \tag{3}$$

Then the empirical lexical salience score is defined as

$$\ell(w) = \text{clip}_{[-b,b]}\left(\log \frac{\widehat{p}_H(w)}{\widehat{p}_A(w)}\right), \tag{4}$$

where $a > 0$ is the smoothing parameter and $b > 0$ is the clipping threshold.

Phrase localization PRESS searches for rewrite locations at the phrase level. For a phrase $z = (w_1, \ldots, w_M)$, PRESS computes the lexical salience score of z by averaging the empirical lexical salience scores over words:

$$L(z) = \frac{1}{|I(z)|} \sum_{j \in I(z)} \ell(w_j), \qquad I(z) = \{j : w_j \in \mathcal{V}\}. \tag{5}$$

Phrases with lower $L(z)$ are considered first because their wording is more strongly associated with AI-edited texts. Among phrases with equal lexical salience scores, we prioritize those with higher estimated formality.

3.3 Candidate Generation and Selection

Once a phrase is selected, the generator uses nearby context to propose several coherent sentence-level rewrites focused on that phrase. Each rewrite replaces the corresponding sentence in the current text to form a candidate text. PRESS ranks each candidate using two signals: the paired humanization prior gain and the surrogate model score. Since the original human-written text is unavailable, the prior gain estimates which candidates are more consistent with human-written patterns, while the surrogate model identifies candidates that are more likely to receive lower AI editing scores. These two signals provide complementary guidance for candidate selection.

Paired humanization prior gain. We compare text representations in a fixed local language model to measure how well a candidate follows the representation difference between original human-written and AI-edited texts captured by the paired humanization prior.

Let $h(z)$ denote the representation of text z, and let $\bar{h}(z) = h(z)/\|h(z)\|_2$ be the normalized representation. For each pair $(U, V) \sim \mathcal{P}_{\text{pair}}$, we consider the representation difference between the original human-written text U and the corresponding AI-edited text V. We define the mean paired representation difference as

$$\mu_* = \mathbb{E}_{(U,V)\sim\mathcal{P}_{\text{pair}}}\big[\bar{h}(U) - \bar{h}(V)\big], \tag{6}$$

and normalize μ_* to obtain the population representation direction

$$d_* = \frac{\mu_*}{\|\mu_*\|_2}. \tag{7}$$

Since $\mathcal{P}_{\text{pair}}$ is unknown in practice, we estimate the population representation direction from the finite paired sample $\mathcal{D}_{\text{pair}} = \{(u_i, v_i)\}_{i=1}^{N}$ by

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\big[\bar{h}(u_i) - \bar{h}(v_i)\big], \qquad d = \frac{\widehat{\mu}}{\|\widehat{\mu}\|_2}, \tag{8}$$

where $\widehat{\mu}$ is the empirical mean paired representation difference and d is the normalized direction.

For a candidate text c, we define the prior gain by combining its movement along the empirical representation direction d with the change in lexical salience:

$$H(x_t, c) = \alpha\langle \bar{h}(c) - \bar{h}(x_t), d\rangle + \beta \tanh\left(\frac{L(c) - L(x_t)}{\sigma}\right). \tag{9}$$

The weights $\alpha, \beta \geq 0$ balance the representation and lexical components. The scale $\sigma > 0$ is determined by the standard deviation of the lexical salience scores, and $\tanh$ limits the magnitude of the lexical term. A larger $H(x_t, c)$ indicates that the candidate is more consistent with the humanization patterns captured by the paired prior.

Surrogate model. We use a surrogate model to provide complementary score guidance for candidate selection. For each paired sample (u_i, v_i), we compute the cosine distance between the original human-written text u_i and the corresponding AI-edited text v_i:

$$D_{\text{CD}}(u_i, v_i) = 1 - \frac{\langle e(u_i), e(v_i)\rangle}{\|e(u_i)\|_2 \|e(v_i)\|_2}, \tag{10}$$

where $e(\cdot)$ is the text encoder used to compute the supervision target. A larger value indicates a larger difference from the original human-written text.

We train a regression head g_ϕ on the paired sample to predict this distance from the AI-edited text alone. For a candidate text c, the surrogate model score is

$$S_{\text{proxy}}(c) = g_\phi(h(c)), \tag{11}$$

where $h(c)$ is the representation of c. The original human-written texts are used only to construct the supervision targets during training and are not required during attack generation. A lower surrogate model score indicates that a candidate is more likely to receive a lower estimate of AI editing.

Combine prior and surrogate rankings. We rank each candidate using both signals. Let $r_s(c)$ denote the rank from lowest to highest surrogate model score, and let $r_h(c)$ denote the rank from highest to lowest prior gain. We combine the two rankings as

$$R(c) = \lambda_s \frac{r_s(c)}{\max(K-1, 1)} + \lambda_h \frac{r_h(c)}{\max(K-1, 1)}, \tag{12}$$

where K is the number of valid candidates and $\lambda_s, \lambda_h \geq 0$ control the relative contributions of the two rankings. Both rankings start at zero for the best candidate. A smaller $R(c)$ therefore indicates a better candidate under the combined humanization and score reduction guidance.

3.4 Quality Checks and Iterative Rewriting

Quality Checks. PRESS uses three checks to control rewrite quality. First, contextual token-matching F1 measures semantic similarity between the text before and after rewriting. Second, content preservation checks protect important information such as numbers, negation, and time references. Third, grammaticality checks identify grammatical and structural problems that require correction.

Iterative Rewriting. With these quality checks in place, PRESS iteratively rewrites the input starting from $x_0 = x$. At each iteration, PRESS first identifies the region of the text to rewrite using lexical salience, then generates candidate rewrites for that location. Candidates that would make the cumulative perturbation rate exceed the budget ϵ or fail the quality checks are removed. The remaining candidates are ranked using the paired humanization prior and surrogate model. The

highest-ranked candidates are post-edited to correct grammatical and structural issues, and are then checked again under the same cumulative budget and quality requirements. The best valid candidate is accepted as x_{t+1}, and the cumulative perturbation rate is updated with all changes introduced so far. If none of the stopping conditions is met, PRESS continues to rewrite the remaining unedited regions while keeping previously edited regions fixed. The process stops when no valid candidate remains, the perturbation budget is exhausted, or the maximum number of iterations is reached. The current text is then returned as the final adversarial text x'. Figure 2 summarizes the overall procedure of PRESS.

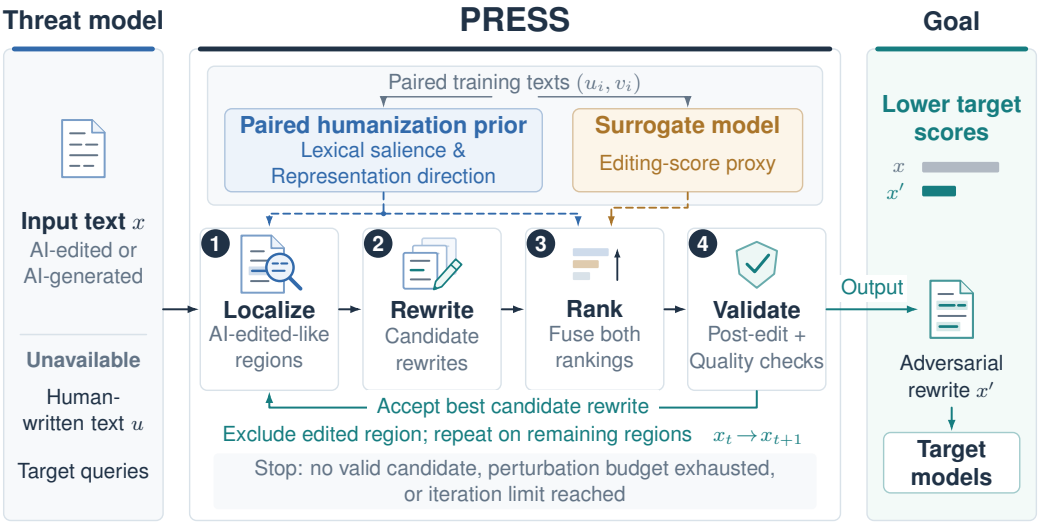


Figure 2: Overview of PRESS. Dashed arrows indicate prior and surrogate guidance; solid arrows show the workflow. The loop repeats on unmodified sentences after each accepted update. Target models evaluate the output x' without guiding its generation.

3.5 Stability of Candidate Rewrite Selection

The ideal candidate selection uses the population representation direction d_* defined in (7). In practice, however, the paired population distribution is difficult to obtain directly, so PRESS uses the empirical direction d estimated from finite paired data in (8). We therefore study whether using d instead of d_* changes the selected candidate.

Let $R(c)$ be the fused ranking defined in (12) computed with the empirical direction d. We define $R_*(c)$ in the same way, but replace d with the population direction d_*. For each candidate c, let $B_t(c)$ be an upper bound on how much the fused ranking can change when d is used instead of d_*, as defined in Appendix A. The following result gives a sufficient condition under which the estimation does not change the selected candidate.

Theorem 1 (Stability of candidate selection)**.** *Let $\hat{c}$ be the candidate selected using the empirical direction d. If $R(c) - R(\hat{c}) > B_t(c) + B_t(\hat{c})$ for every $c \neq \hat{c}$, then $\hat{c}$ is also the unique candidate selected using the population direction d_*.*

The theorem states that when the candidate rankings are sufficiently well separated, using the estimated direction d selects the same candidate as using the population direction d_*. When the rankings are less clearly separated, the selected candidate is still guaranteed to remain close to the best candidate under d_*. Detailed bounds and proofs are provided in Appendix A.

4 Experiments

We evaluate PRESS on three evaluation splits and five evaluation targets. We first test whether the same adversarial texts reduce target scores across models and transfer to texts produced by other

Table 1: Score reductions across evaluation splits and targets.

Data split	Target	Original	Adversarial	Mean Δ_T	$P_{\downarrow}$	$P_{10\%}$	$P_{20\%}$
Test	Llama	0.6610	0.5886	0.0723	75.69%	40.95%	24.56%
	RoBERTa	0.7135	0.6930	0.0205	74.31%	16.61%	9.49%
	DeBERTa-large	0.7208	0.6830	0.0377	78.69%	22.88%	11.87%
	ModernBERT-large	0.7295	0.6981	0.0314	81.70%	19.35%	7.70%
	Fast-DetectGPT	2.6270	2.2619	0.3651	76.77%	54.21%	31.46%
Test_llama	Llama	0.7054	0.6398	0.0656	80.33%	38.17%	21.97%
	RoBERTa	0.7108	0.6776	0.0332	75.86%	22.81%	14.77%
	DeBERTa-large	0.7075	0.6533	0.0542	79.59%	30.72%	17.61%
	ModernBERT-large	0.7077	0.6582	0.0495	78.95%	28.32%	13.32%
	Fast-DetectGPT	5.4535	4.3973	1.0561	89.05%	69.56%	43.80%
Test_enron	Llama	0.7213	0.6914	0.0299	73.65%	20.48%	8.01%
	RoBERTa	0.5165	0.4556	0.0609	63.06%	35.98%	25.39%
	DeBERTa-large	0.6280	0.5562	0.0718	72.18%	35.76%	22.19%
	ModernBERT-large	0.5947	0.5279	0.0668	72.56%	38.41%	23.06%
	Fast-DetectGPT	1.4120	1.1395	0.2726	67.70%	57.03%	41.86%

LLMs and to a different text domain. We then examine whether lower AI editing scores require the adversarial texts to move closer to the corresponding original human-written texts. Next, we study the amount and quality of rewriting and the effect of the perturbation budget. Finally, we compare PRESS with T5 paraphrasing to examine whether PRESS achieves stronger score reduction than a general paraphrasing baseline.

4.1 Experimental Setup

Data. We use four splits of the EditLens dataset (Thai et al., 2026): Train, Test, Test_llama, and Test_enron.[1] The dataset contains AI-edited texts, AI-generated texts, and original human-written texts. Train and Test share the same text domains and AI generation sources. We use 22,265 pairs of original human-written texts and the corresponding AI-edited texts from Train to estimate the paired humanization prior and train the surrogate model, while Test contains 4,077 texts for in-distribution evaluation. Test_llama keeps the same text domains but changes the AI generation source, with 3,919 texts used to evaluate generalization to texts produced by a different LLM. Test_enron contains 4,155 texts from a different domain and is used to evaluate generalization across text domains. Additional dataset details are provided in Appendix B.1.

Evaluation targets. We evaluate two EditLens models based on RoBERTa-large (RoBERTa) and Llama-3.2-3B (Llama) (Thai et al., 2026), together with two additional models based on DeBERTa-v3-large (DeBERTa-large) and ModernBERT-large (ModernBERT-large) that predict the extent of AI editing. We also evaluate Fast-DetectGPT (Bao et al., 2024) and report its raw Fast-DetectGPT criterion. Details and validation results are provided in Appendix B.2 and Appendix C.

PRESS settings. Gemma 3 4B IT is used for candidate rewriting and as the surrogate backbone, while Qwen2.5-7B-Instruct is used for post-editing. We set the word-level perturbation budget to $\epsilon = 0.50$, use equal ranking weights $\lambda_s = \lambda_h = 1$, and require a contextual token-matching F1 of at least $\tau = 0.70$. Further implementation details are provided in Appendix B.3.

4.2 Transfer Across Models and Domains

Let $S_T(x)$ denote the target value of the original input and $S_T(x')$ the value of the adversarial text. We report the mean reduction $\Delta_T = S_T(x) - S_T(x')$ and

$$P_{\downarrow} = \Pr(\Delta_T > 0), \qquad P_{q\%} = \Pr\Big(\Delta_T \geq \frac{q}{100} S_T(x) \mid S_T(x) > 0\Big), \tag{13}$$

[1] https://huggingface.co/datasets/pangram/editlens_iclr

Table 2: Word-level perturbation rate and contextual token-matching similarity.

Measure	Test	Test_llama	Test_enron
Mean perturbation rate	17.42%	21.86%	21.45%
Mean token-matching F1	0.844	0.836	0.849

Table 3: Score reduction and change in cosine distance on AI-edited texts.

Measure	Test	Test_llama	Test_enron
Llama Δ_T	0.0331	0.0375	0.0268
RoBERTa Δ_T	0.0258	0.0364	0.0105
DeBERTa-large Δ_T	0.0328	0.0366	0.0224
ModernBERT-large Δ_T	0.0279	0.0239	0.0164
Cosine-distance change δ_{CD}	+0.00232	+0.00214	+0.00640

where $P_{\downarrow}$ is the proportion of inputs with lower target values and $P_{q\%}$ is the proportion with a relative reduction of at least $q\%$.

As shown in Table 1, across the four models that quantify the extent of AI editing, PRESS consistently lowers AI editing scores on all three evaluation splits. On Test, PRESS achieves a mean reduction of 0.0723 for Llama, and 75.69% of the inputs receive lower scores. On Test_llama, the mean reductions are 0.0656 for Llama, 0.0542 for DeBERTa-large, and 0.0495 for ModernBERT-large. About 80% of the inputs receive lower scores under these three models. The attack also transfers to Test_enron. The mean reductions reach 0.0718 for DeBERTa-large and 0.0668 for ModernBERT-large. More than 35% of the inputs receive reductions of at least 10% under both models. The same adversarial texts also lower the Fast-DetectGPT criterion. The criterion decreases for 76.77% of Test inputs, 89.05% of Test_llama inputs, and 67.70% of Test_enron inputs, which shows that PRESS also transfers to an AI detector with a different scoring objective.

As shown in Table 2, the score reductions are achieved with mean perturbation rates of 17.42–21.86% and contextual token-matching F1 scores of 0.836–0.849, indicating that PRESS maintains high semantic similarity with moderate text changes.

4.3 Score Reduction and Human-Text Distance

For AI-edited inputs, we next test whether the score reduction occurs only because the adversarial text moves closer to the corresponding original human-written text. For an original input x with original human-written text u, we define

$$\delta_{\mathrm{CD}} = D_{\mathrm{CD}}(u, x') - D_{\mathrm{CD}}(u, x). \tag{14}$$

A positive value means that the adversarial text is farther from the original human-written text. As shown in Table 3, all four models give lower mean AI editing scores on all three splits, while the mean cosine distance increases. Thus, lower AI editing scores do not require the adversarial text to move closer to the original human-written text, which suggests that the score may not reliably reflect the actual extent of AI editing under adversarial rewriting. Additional analyses are reported in Appendix C.

4.4 Perturbation Amount, Quality, and Budget Sensitivity

We further vary the perturbation budget from $\epsilon = 0.10$ to $\epsilon = 0.50$ on 1,000 randomly sampled Test inputs. PRESS reduces both Llama and RoBERTa scores at every budget. With a 10% budget, the mean Llama reduction is 0.0322 with only 7.90% of words changed on average. With a 20% budget, the mean Llama reduction reaches 0.0605, about 80% of the reduction under the 50% budget, while the mean perturbation rate is 12.78%. The gains become smaller beyond a 30% budget, showing that much of the score reduction is obtained before the largest budget is used.

Table 4: Score reductions under different word-level perturbation budgets.

Budget ϵ	Mean perturbation rate	Token-matching F1	Llama Δ_T	RoBERTa Δ_T
0.10	7.90%	0.869	0.0322	0.0079
0.20	12.78%	0.849	0.0605	0.0144
0.30	15.36%	0.844	0.0658	0.0192
0.40	16.31%	0.843	0.0707	0.0194
0.50	17.14%	0.842	0.0760	0.0218

Table 5: Mean score reduction for T5 paraphrasing and PRESS under the same word-level perturbation budget of $\epsilon = 0.50$. Higher values indicate larger score reductions.

	Test		Test_llama		Test_enron	
Target	T5	PRESS	T5	PRESS	T5	PRESS
Llama	0.0339	**0.0723**	0.0080	**0.0656**	0.0141	**0.0299**
RoBERTa	0.0197	**0.0205**	0.0178	**0.0332**	0.0427	**0.0609**
DeBERTa-large	0.0231	**0.0377**	0.0185	**0.0542**	0.0426	**0.0718**
ModernBERT-large	0.0228	**0.0314**	0.0211	**0.0495**	0.0475	**0.0668**
Fast-DetectGPT	0.0753	**0.3651**	0.2271	**1.0561**	0.1011	**0.2726**

4.5 Comparison with T5 Paraphrasing

We compare PRESS with the released T5 paraphraser trained on ChatGPT paraphrases (Vorobev & Kuznetsov, 2023) under the same word-level perturbation budget of $\epsilon = 0.50$. PRESS gives larger mean reductions than T5 on all five targets across all three evaluation splits. On Test, the Llama reduction increases from 0.0339 to 0.0723, while the Fast-DetectGPT reduction increases from 0.0753 to 0.3651. On Test_llama, the Llama reduction increases from 0.0080 to 0.0656, and the Fast-DetectGPT reduction increases from 0.2271 to 1.0561. On Test_enron, the DeBERTa-large reduction increases from 0.0426 to 0.0718, while the Fast-DetectGPT reduction increases from 0.1011 to 0.2726. These results show that PRESS provides stronger score reduction than T5 paraphrasing across models and evaluation splits.

5 Conclusion

We test the robustness of AI editing scores under adversarial rewriting when the original human-written text paired with the input is unavailable. To this end, we develop PRESS, an adversarial rewriting attack that reduces AI editing scores while preserving meaning and controlling the amount of text modification. PRESS constructs a paired humanization prior from original human-written and AI-edited text pairs to identify regions that are more characteristic of AI-edited writing and to guide candidate selection toward more human-like lexical and representation patterns. A surrogate model provides complementary guidance by estimating the AI editing scores of candidate rewrites. By combining these two signals, PRESS can select effective rewrites without querying the target model or using the original human-written text paired with the current input during attack generation. Our theoretical analysis further establishes the stability of candidate selection when the humanization prior is estimated from finite paired data, showing that the estimated prior can recover the same candidate as the population prior under sufficient ranking separation and remains close to the population optimum otherwise.

Experiments show consistent score reductions across four models that quantify the extent of AI editing, including transfer to texts produced by another language model and to another domain. The same adversarial texts also reduce the Fast-DetectGPT criterion, showing transfer to an AI detector with a different scoring objective. For AI-edited texts, lower scores do not require the adversarial text to move closer to the original human-written text. These findings raise concerns about the reliability of AI editing scores as measures of the extent of AI editing when the original human-written text is unavailable.

REFERENCES

Answer.AI and LightOn. ModernBERT-large. Hugging Face model card, 2024. URL https://huggingface.co/answerdotai/ModernBERT-large.

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 *Proceedings of the 12th International Conference on Learning Representations*, pp. 24814–24836, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6b8c6f846c3575e1d1ad496abea28826-Abstract-Conference.html.

Google. Gemma 3 4B IT. Hugging Face model card, 2025. URL https://huggingface.co/google/gemma-3-4b-it.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17519–17537. PMLR, 2024. URL https://proceedings.mlr.press/v235/hans24a.html.

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

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, volume 34, pp. 8018–8025. AAAI Press, 2020. doi: 10.1609/aaai.v34i05.6311. URL https://ojs.aaai.org/index.php/AAAI/article/view/6311.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27469–27500, 2023. doi: 10.52202/075280-1195. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/575c450013d0e99e4b0ecf82bd1afaa4-Abstract-Conference.html.

Eric Lei, Hsiang Hsu, and Chun-Fu Chen. PaLD: Detection of text partially written by large language models. In *Proceedings of the 13th International Conference on Learning Representations*, pp. 86593–86611, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d799b1d6a5e43546e67e7afdeffc067d-Abstract-Conference.html.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6193–6202. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.500. URL https://aclanthology.org/2020.emnlp-main.500/.

Yang Li, Qiang Sheng, Zhengjia Wang, Yehan Yang, Danding Wang, and Juan Cao. Beyond the final actor: Modeling the dual roles of creator and editor for fine-grained LLM-generated text detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5188–5203. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.235. URL https://aclanthology.org/2026.acl-long.235/.

Microsoft. DeBERTa-v3-large. Hugging Face model card, 2021. URL https://huggingface.co/microsoft/deberta-v3-large.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24950–24962. PMLR, 2023. URL https://proceedings.mlr.press/v202/mitchell23a.html.

Charlotte Nicks, Eric Mitchell, Rafael Rafailov, Archit Sharma, Christopher D. Manning, Chelsea Finn, and Stefano Ermon. Language model detectors are easily optimized against. In *Proceedings of the 12th International Conference on Learning Representations*, pp. 7807–7826, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1f9f07df0992ce21698d800eaa891bd8-Abstract-Conference.html.

Pangram. EditLens Llama-3.2-3B model checkpoint. Hugging Face model card, 2026a. URL https://huggingface.co/pangram/editlens_Llama-3.2-3B.

Pangram. EditLens RoBERTa-large model checkpoint. Hugging Face model card, 2026b. URL https://huggingface.co/pangram/editlens_roberta-large.

Qwen Team. Qwen2.5-7B-Instruct-GGUF. Hugging Face model card, 2024. URL https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-GGUF.

Suraj Ranganath and Atharv Ramesh. StealthRL: Reinforcement learning paraphrase attacks for multi-detector evasion of AI-text detectors. *arXiv preprint arXiv:2602.08934*, 2026. URL https://github.com/suraj-ranganath/StealthRL.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-generated text be reliably detected? Stress testing AI text detectors under various attacks. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=OOgsAZdFOt.

Shoumik Saha and Soheil Feizi. Almost AI, almost human: The challenge of detecting AI-polished writing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25414–25431. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1303. URL https://aclanthology.org/2025.findings-acl.1303/.

Sentence-Transformers. all-MiniLM-L6-v2. Hugging Face model card, 2021. URL https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. Red teaming language model detectors with language models. *Transactions of the Association for Computational Linguistics*, 12:174–189, 2024. doi: 10.1162/tacl_a_00639. URL https://aclanthology.org/2024.tacl-1.10/.

Katherine Thai, Bradley Emi, Elyas Masrour, and Mohit Iyyer. EditLens: Quantifying the extent of AI editing in text. In *Proceedings of the 14th International Conference on Learning Representations*, pp. 122997–123024, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c7d936e696f57468b6b80a7b1463178c-Abstract-Conference.html.

Vladimir Vorobev and Maxim Kuznetsov. A paraphrasing model based on ChatGPT paraphrases. Hugging Face model card, 2023. URL https://huggingface.co/humarin/chatgpt_paraphraser_on_T5_base.

Yixuan Even Xu, Ziqian Zhong, Aditi Raghunathan, Fei Fang, and J. Zico Kolter. Base models look human to AI detectors. *arXiv preprint arXiv:2605.19516*, 2026. URL https://github.com/YixuanEvenXu/humanization-by-iterative-paraphrasing.

Lingyi Yang, Feng Jiang, and Haizhou Li. Is ChatGPT involved in texts? Measure the Polish Ratio to detect ChatGPT-generated text. *APSIPA Transactions on Signal and Information Processing*, 13 (2):1–19, 2024. doi: 10.1561/116.00000250. URL https://www.emerald.com/atsip/article-lookup/doi/10.1561/116.00000250.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. 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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.29. URL https://aclanthology.org/2024.findings-naacl.29/.

A PROOFS FOR CANDIDATE SELECTION STABILITY

We provide the theoretical results used to justify candidate selection with the empirical representation direction. The analysis first controls the error between the empirical and population directions, then translates this error into a bound on the fused ranking, and finally proves the candidate selection result stated in Theorem 1.

A.1 FINITE-SAMPLE STABILITY OF THE REPRESENTATION DIRECTION

For the concentration argument, we group paired samples that share the same original human-written text. Suppose there are G independent groups. Group g contains one original human-written text u_g and n_g corresponding AI-edited texts $v_{g1}, \dots, v_{gn_g}$. Let $N = \sum_{g=1}^{G} n_g$ and $w_g = n_g/N$. Dependence among AI-edited texts within the same group is allowed, while different groups are assumed to be independent.

Define the average representation difference in group g by

$$Q_g = \frac{1}{n_g} \sum_{j=1}^{n_g} \big[\bar{h}(u_g) - \bar{h}(v_{gj})\big].$$

Then the empirical mean paired representation difference can be written as $\widehat{\mu} = \sum_{g=1}^{G} w_g Q_g$, and its population counterpart is $\mu_* = \sum_{g=1}^{G} w_g \mathbb{E}Q_g$. Let $\nu_g = \mathbb{E}\|Q_g - \mathbb{E}Q_g\|_2^2$, and assume $\|Q_g\|_2 \le B_q$ for every group.

Lemma 2 (Finite-sample stability of the representation direction)**.** *For any $0 < \delta < 1$, define*

$$r_\delta = \sqrt{\sum_{g=1}^{G} w_g^2 \nu_g} + B_q \sqrt{2\log(1/\delta) \sum_{g=1}^{G} w_g^2}.$$

With probability at least $1-\delta$, $\|\widehat{\mu} - \mu_*\|_2 \le r_\delta$. If $\|\widehat{\mu}\|_2 > r_\delta$, then the population direction $d_* = \mu_*/\|\mu_*\|_2$ is well defined and

$$\|d - d_*\|_2 \le \eta, \qquad \eta = \min\left\{2, \frac{2r_\delta}{\|\widehat{\mu}\|_2}\right\},$$

where $d = \widehat{\mu}/\|\widehat{\mu}\|_2$ is the empirical direction in (8)*.*

Proof. Let $X_g = Q_g - \mathbb{E}Q_g$. Since the groups are independent and $\mathbb{E}X_g = 0$,

$$\mathbb{E}\|\widehat{\mu} - \mu_*\|_2^2 = \sum_{g=1}^{G} w_g^2 \nu_g.$$

Jensen's inequality therefore gives $\mathbb{E}\|\widehat{\mu} - \mu_*\|_2 \le \sqrt{\sum_g w_g^2 \nu_g}$.

Replacing one group value Q_g by another admissible value Q_g' changes $\|\widehat{\mu} - \mu_*\|_2$ by at most $2B_q w_g$. McDiarmid's inequality then implies that, with probability at least $1-\delta$, the deviation from its expectation is at most $B_q\sqrt{2\log(1/\delta)\sum_g w_g^2}$. Combining the expectation and concentration bounds yields $\|\widehat{\mu} - \mu_*\|_2 \le r_\delta$.

Further, if $\|\widehat{\mu}\|_2 > r_\delta$, then $\|\mu_*\|_2 \ge \|\widehat{\mu}\|_2 - r_\delta > 0$, so d_* is well defined. Adding and subtracting $\mu_*/\|\widehat{\mu}\|_2$ and using the reverse triangle inequality gives

$$\|d - d_*\|_2 \le 2\|\widehat{\mu} - \mu_*\|_2/\|\widehat{\mu}\|_2 \le 2r_\delta/\|\widehat{\mu}\|_2.$$

Since d and d_* are unit vectors, their distance is also at most 2. Combining the two bounds gives the stated value of η. □

A.2 Stability of the Fused Ranking

Let $H_*(x_t, c)$ denote the prior gain in (9) with d_* replacing the empirical direction d. Likewise, let $R_*(c)$ denote the fused ranking in (12) obtained from H_*, while the lexical salience, surrogate model ranking, candidate set, and tie rule remain unchanged.

For a realized candidate set $\mathcal{C}_t$, define the number of uncertain prior comparisons involving candidate c as

$$a_t(c) = \#\left\{c' \in \mathcal{C}_t \setminus \{c\} : |H(x_t, c) - H(x_t, c')| \le \alpha\eta \|\bar{h}(c) - \bar{h}(c')\|_2\right\},$$

and define

$$B_t(c) = \frac{\lambda_h a_t(c)}{\max\{|\mathcal{C}_t| - 1, 1\}}.$$

Then $B_t(c)$ is an upper bound on how much the fused ranking of candidate c can change when the empirical direction d is used instead of the population direction d_*.

Lemma 3 (Stability of the fused ranking)**.** *On the event in Lemma 2, every candidate $c \in \mathcal{C}_t$ satisfies*

$$|R(c) - R_*(c)| \le B_t(c).$$

Proof. Consider two candidates c and c'. The lexical part of the prior gain is identical under H and H_*, so only the representation direction can change their relative prior gain. By Cauchy–Schwarz and Lemma 2,

$$|[H(x_t, c) - H(x_t, c')] - [H_*(x_t, c) - H_*(x_t, c')]| \le \alpha\eta \|\bar{h}(c) - \bar{h}(c')\|_2.$$

Therefore, any empirical comparison outside the uncertainty condition in the definition of $a_t(c)$ keeps the same order under H_*. At most $a_t(c)$ pairwise comparisons involving c can change, so the prior rank satisfies $|r_h(c) - r_{h,\star}(c)| \le a_t(c)$.

The surrogate model does not use d, so its ranking is unchanged. Applying the fused ranking rule in (12) gives $|R(c) - R_*(c)| \le B_t(c)$.

□

A.3 Population Ranking Gap

The previous lemma controls the ranking of every candidate individually. It also gives a bound when the empirically selected candidate is not guaranteed to be exactly the population choice.

Lemma 4 (Population ranking gap)**.** *Let $\hat{c}$ minimize the empirical fused ranking R over $\mathcal{C}_t$. Then*

$$0 \le R_*(\hat{c}) - \min_{c \in \mathcal{C}_t} R_*(c) \le B_t(\hat{c}) + \max_{c \in \mathcal{C}_t} B_t(c).$$

Proof. Let c_* minimize R_* over $\mathcal{C}_t$. Since $R(\hat{c}) \le R(c_*)$ and Lemma 3 gives $|R(c) - R_*(c)| \le B_t(c)$ for every candidate,

$$R_*(\hat{c}) - R_*(c_*) \le B_t(\hat{c}) + B_t(c_*) \le B_t(\hat{c}) + \max_{c \in \mathcal{C}_t} B_t(c).$$

The lower bound follows because c_* minimizes R_*. □

A.4 Proof of Theorem 1

Proof. Let $\hat{c}$ be the candidate selected using the empirical fused ranking R. Lemma 3 gives $R_*(c) \ge R(c) - B_t(c)$ for every candidate and $R_*(\hat{c}) \le R(\hat{c}) + B_t(\hat{c})$. Hence, for any $c \ne \hat{c}$,

$$R_*(c) - R_*(\hat{c}) \ge R(c) - R(\hat{c}) - B_t(c) - B_t(\hat{c}).$$

Under the separation condition in Theorem 1, the right-hand side is positive for every competing candidate. Therefore $R_*(\hat{c}) < R_*(c)$ for all $c \ne \hat{c}$, so $\hat{c}$ is also the unique candidate selected using the population direction d_*. □

A.5 Effect of Pairing on the Representation Estimate

The paired construction can reduce the variability of the representation difference when the original human-written text and its corresponding AI-edited text share common variation. Let $U = \bar{h}(u)$ and $V = \bar{h}(v)$ for a matched pair. Define $\Gamma = \mathbb{E}\langle U - \mathbb{E}U, V - \mathbb{E}V\rangle$. Then

$$\mathbb{E}\|(U - V) - \mathbb{E}(U - V)\|_2^2 = \mathbb{E}\|U - \mathbb{E}U\|_2^2 + \mathbb{E}\|V - \mathbb{E}V\|_2^2 - 2\Gamma.$$

When $\Gamma > 0$, the covariance term reduces the variance of the paired difference and can improve the stability of the estimated representation direction. Pairing alone does not guarantee a smaller variance. The benefit comes from positive dependence within matched pairs.

B Experimental Details

This section gives more details about the data, models, candidate processing, acceptance checks, and evaluation metrics.

B.1 Data Composition

Table 6 summarizes the data used in the experiments. We use 22,265 valid pairs of original human-written and AI-edited texts from Train to build the paired humanization prior and train the surrogate model. Each pair contains a nonempty human source and the corresponding edited text and contains at least 75 words.

Table 6: Dataset composition.

Split	Text type	Text domain	AI model	Count
Train	Human-written	Reviews, creative writing, educational web articles, and news	—	18,810
	AI-edited	Same domains as Test	GPT-4.1, Claude 4 Sonnet, and Gemini 2.5 Flash	22,381
	AI-generated	Same domains as Test	GPT-4.1, Claude 4 Sonnet, and Gemini 2.5 Flash	18,809
Test	AI-edited	Amazon and Google reviews, Reddit Writing Prompts, FineWeb-EDU, XSum, and CNN/DailyMail	GPT-4.1, Claude 4 Sonnet, and Gemini 2.5 Flash	2,031
	AI-generated	Same domains	GPT-4.1, Claude 4 Sonnet, and Gemini 2.5 Flash	2,046
Test_llama	AI-edited	Same domains as Test	Llama-3.3-70B-Instruct-Turbo	1,881
	AI-generated	Same domains as Test	Llama-3.3-70B-Instruct-Turbo	2,038
Shared controls	Human-written	Same domains as Test	—	2,038
Test_enron	AI-edited or generated	Enron emails	GPT-4.1, Claude 4 Sonnet, and Gemini 2.5 Flash	4,155
	Human-written controls	Enron emails	—	1,992

B.2 Evaluation Scorers

For the two released EditLens models (Pangram, 2026b;a), the continuous editing extent score is computed from the four editing levels as

$$S_T(x) = \sum_{k=0}^{B-1} \frac{k}{B-1}\, p_T(k \mid x), \qquad B = 4, \tag{15}$$

where $p_T(k \mid x)$ is the predicted probability of editing level k. The score ranges from 0 to 1, and a higher value indicates greater predicted editing extent.

We also train two additional editing extent scorers using DeBERTa-v3-large (Microsoft, 2021) and ModernBERT-large (Answer.AI & LightOn, 2024) as pretrained backbones. We use the same four editing level labels as EditLens and convert the predicted class probabilities to continuous scores using (15). Their validation results are reported in Appendix C.

Fast-DetectGPT is evaluated separately because it measures evidence of machine-generated text rather than editing extent. We use its raw conditional probability curvature criterion. A higher Fast-DetectGPT score indicates stronger evidence of machine generation. Its numerical scale is therefore different from that of the editing extent scores.

All scorers evaluate the same adversarial texts, and PRESS does not perform target-specific optimization. The released EditLens Llama model uses an input limit of 1,024 tokens, while the released RoBERTa model uses 512 tokens.

B.3 Attack Configuration Details

Table 7 lists the main models used to generate adversarial texts. Gemma 3 4B IT (Google, 2025) is used for candidate generation and also provides the frozen backbone of the surrogate model. Qwen2.5-7B-Instruct (Qwen Team, 2024) is used only for grammatical repair. all-MiniLM-L6-v2 (Sentence-Transformers, 2021) is used for local semantic similarity.

Table 7: Models used during attack generation.

Role	Checkpoint	Setting
Candidate generation	`google/gemma-3-4b-it`	NF4
Surrogate backbone	`google/gemma-3-4b-it`	NF4
Sentence repair	`Qwen/Qwen2.5-7B-Instruct-GGUF`	Q4_K_M
Semantic similarity	`sentence-transformers/all-MiniLM-L6-v2`	Full precision

Prior and surrogate settings. The paired humanization prior is built from the 22,265 training pairs described in Appendix B.1. For the lexical prior, words with a combined frequency below 12 are removed. We use smoothing parameter $a = 1$ and clipping parameter $b = 4$. The relative prior gain uses $\alpha = 1$ and $\beta = 0.25$. The scale σ is the standard deviation of the retained lexical weights.

The surrogate uses the Gemma 3 4B IT backbone with a maximum input length of 512 tokens. We take the mean of the nonpadding token representations from the penultimate hidden layer. The resulting 2,560-dimensional representation is passed to a regression head with dimensions $2560 \to 256 \to 128 \to 1$. ReLU is used after the first two linear layers, with dropout of 0.2 after the first activation. The Gemma backbone remains fixed. The scalar output of the regression head is used for candidate ranking.

Candidate generation and ranking. Candidate regions are selected using the lexical part of the paired humanization prior. Previously modified sentences are not selected again. For each region, Gemma receives the selected sentence, the previous and next sentences when available, and the remaining edit budget. Generation uses temperature 0.70 and nucleus sampling with $p = 0.92$. We generate up to eight distinct sentence candidates for each selected region. Each generation call returns up to four candidates, with at most six calls for one region. The prompt and reserved output together use at most 2,048 tokens, and the output length is limited to 224 tokens.

Candidates that pass the initial edit and content checks are ranked using the paired humanization prior and surrogate model. We set $\lambda_s = \lambda_h = 1$. The top three candidates are sent to the repair stage. After repair, the prior and surrogate scores are recomputed before the final candidate is selected.

Repair and acceptance. Qwen2.5-7B-Instruct is run with the Q4_K_M GGUF checkpoint through `llama.cpp`. Repair uses temperature 0.3, nucleus sampling with $p = 0.9$, and a 2,048-token context window. Each shortlisted candidate can be repaired for at most three rounds.

After repair, the complete candidate document is checked again. The protected information checks cover numbers and units, selected names and identifiers, negation, time references, and document formatting. Local semantic F1 is computed with all-MiniLM-L6-v2 between the original sentence and the accepted rewrite, and must be at least $\tau = 0.70$. Candidates with unresolved grammatical or structural errors are rejected. Reported semantic F1 values are averaged over accepted edits within each changed document and then over documents with available measurements.

Edit and search budgets. The main experiments use a cumulative document edit budget of $\epsilon = 0.50$. The word edit rate compares the original input x with the current adversarial text x'. We split both texts on whitespace and align them using `SequenceMatcher`. For each aligned block b, let m_b and n_b denote the numbers of words in x and x', respectively. The edit rate is defined as

$$E(x, x') = \frac{\sum_{b \in \mathcal{B}(x,x')} \mathrm{cost}(b)}{\max(|x|_{\mathrm{word}}, 1)}, \tag{16}$$

where an equal block has cost 0, a replacement has cost $\max(m_b, n_b)$, an insertion has cost n_b, and a deletion has cost m_b. The edit budget includes all changes from both candidate rewriting and grammatical repair. Unchanged outputs remain in the score and edit rate analyses.

Each successful iteration accepts at most one sentence replacement. The initial surrogate score selects one of three search levels, which allow at most four, six, or eight accepted sentence replacements, respectively. Table 8 gives the corresponding computation limits.

Table 8: Search budget used for attack generation.

Initial surrogate signal	Max. sentence replacements	Budget multiplier
$[0, 0.90)$	4	1.0
$[0.90, 0.99)$	6	1.5
$[0.99, 1]$	8	2.0

The base limits are 144 Gemma generation calls, 18,432 generated tokens, and 220 semantic similarity checks per input. The multiplier in Table 8 scales these limits. The maximum numbers of Qwen repair calls and surrogate evaluations are fixed at 36 and 800, respectively. The search stops when no valid candidate is found or when an edit, iteration, or computation limit is reached.

C Additional Experimental Results

The following results provide additional checks for the evaluation scorers, transfer across text domains and input types, editing cost, and source distance.

Validation of the additional editing extent scorers. Table 9 compares the two trained scorers with the released EditLens models. ModernBERT-large reaches an accuracy of 0.8383 and a macro-F1 of 0.7122, close to those of EditLens Llama. DeBERTa-large reaches an accuracy of 0.8275 and a macro-F1 of 0.6870. Both trained scorers exceed EditLens RoBERTa in accuracy and macro-F1.

Transfer across text domains. Table 10 reports mean score reductions for the released EditLens scorers across the source text domains. Positive reductions hold for reviews, creative writing, educational web articles, and news on both Test and Test_llama. The pattern therefore remains across text types as well as generation models. Test uses the 4,076 AI inputs with paired EditLens scores, and Test_llama uses all 3,919 AI inputs.

Table 9: Classification performance of the editing extent scorers.

Model	Acc.	Macro-F1	$F1_0$	$F1_1$	$F1_2$	$F1_3$
EditLens Llama	0.8454	0.7202	0.9243	0.6354	0.3742	0.9470
ModernBERT-large	0.8383	0.7122	0.9257	0.6637	0.3273	0.9322
DeBERTa-large	0.8275	0.6870	0.9271	0.6404	0.2611	0.9192
EditLens RoBERTa	0.8113	0.6076	0.9171	0.5722	0.0429	0.8983

Table 10: Mean score reductions across source text domains.

	Test			Test_llama		
Text domain	n	Llama	RoBERTa	n	Llama	RoBERTa
Amazon reviews	569	0.0852	0.0347	526	0.0628	0.0287
Google reviews	418	0.0990	0.0384	391	0.0557	0.0425
Reddit Writing Prompts	1,087	0.0946	0.0246	1,038	0.0925	0.0574
FineWeb-EDU	950	0.0626	0.0176	1,008	0.0710	0.0295
News	1,052	0.0406	0.0039	956	0.0365	0.0095

Results by input type. Table 11 separates AI-edited texts, AI-generated texts, and human controls. All four editing extent scorers have positive mean reductions for every AI group. Human controls move in the opposite direction on Test and Test_llama and also on Test_enron. The contrast shows that rewriting does not produce a uniform downward shift for arbitrary text.

Table 11: Mean score reduction by input type. Negative values denote score increases.

Split	Text type	Llama	RoBERTa	DeBERTa-large	ModernBERT-large
Test	AI-edited	0.0331	0.0258	0.0328	0.0279
Test	AI-generated	0.1113	0.0152	0.0427	0.0348
Test/Test_llama	Human controls	−0.0293	−0.0334	−0.0551	−0.0475
Test_llama	AI-edited	0.0375	0.0364	0.0366	0.0239
Test_llama	AI-generated	0.0917	0.0303	0.0705	0.0731
Test_enron	AI-edited	0.0217	0.0102	0.0180	0.0154
Test_enron	AI-generated	0.0401	0.1242	0.1390	0.1311
Test_enron	Human controls	−0.0462	−0.0295	−0.0665	−0.0507

Editing cost and local similarity. Table 12 gives additional editing statistics. Most outputs remain well below the 50% cap. On Test and Test_llama, the median edit rates are 14.81% and 20.49%, and 66.69% and 48.92% of outputs change at most 20% of words. The median local token F1 values are 0.851 and 0.841.

Paired score and source distance analysis. For an AI-edited text, the event $\Delta_T > 0$ and $\delta_{\mathrm{CD}} \geq 0$ means that the editing extent score decreases without the adversarial text moving closer to the human source. Table 13 shows that the event occurs for about 29%–36% of inputs across all four editing extent scorers on Test and Test_llama. Even when a score reduction greater than 0.05 is required, the proportion of inputs satisfying the event remains between 10.9% and 15.1%. The corresponding Llama proportion is 35.7% on Test_enron.

We also map source distance to the four EditLens editing levels:

$$q(D) = \begin{cases} 0, & D \leq 0.03, \\ 1/3, & 0.03 < D < 0.09, \\ 2/3, & 0.09 \leq D < 0.15, \\ 1, & D \geq 0.15. \end{cases} \tag{17}$$

Table 12: Additional editing statistics.

Measure	Test	Test_llama	Test_enron
Mean edit rate	17.42%	21.86%	21.45%
Outputs with edit rate $\leq 30\%$	85.26%	72.03%	71.93%
Mean local token F1	0.844	0.836	0.849

Table 13: Lower editing extent scores without a decrease in source distance. Intervals are paired 95% bootstrap intervals.

Split	Scorer	$\Delta_T > 0$	$\Delta_T > 0.05$
Test	Llama	32.0% [29.9, 34.1]	14.6% [13.0, 16.1]
	RoBERTa	32.2% [30.2, 34.2]	11.3% [9.9, 12.7]
	DeBERTa-large	33.1% [31.1, 35.1]	12.9% [11.4, 14.3]
	ModernBERT-large	36.0% [33.9, 38.1]	12.5% [11.1, 13.9]
Test_llama	Llama	30.9% [28.8, 33.1]	15.1% [13.5, 16.7]
	RoBERTa	29.0% [26.9, 31.0]	14.0% [12.4, 15.6]
	DeBERTa-large	30.2% [28.1, 32.2]	14.1% [12.6, 15.8]
	ModernBERT-large	29.7% [27.7, 31.7]	10.9% [9.5, 12.3]

For $z \in \{x, x'\}$, the reference error is $|S_T(z) - q(D_{\mathrm{CD}}(u, z))|$. Table 14 shows that mean absolute error increases after the attack for all four editing extent scorers on both test sets. All paired 95% intervals remain above zero.

Table 14: Change in error against the source-based editing level.

Split	Scorer	Before	After	Change [95% CI]
Test	Llama	0.1788	0.2068	0.0280 [0.0227, 0.0334]
	RoBERTa	0.1939	0.2039	0.0099 [0.0044, 0.0155]
	DeBERTa-large	0.1844	0.1941	0.0097 [0.0043, 0.0150]
	ModernBERT-large	0.1766	0.1864	0.0098 [0.0050, 0.0145]
Test_llama	Llama	0.1877	0.2117	0.0240 [0.0181, 0.0297]
	RoBERTa	0.2229	0.2412	0.0183 [0.0117, 0.0248]
	DeBERTa-large	0.1954	0.2077	0.0123 [0.0065, 0.0185]
	ModernBERT-large	0.1901	0.2047	0.0146 [0.0094, 0.0201]

The result remains when the source-based editing level itself does not change. The restriction keeps 1,752 Test inputs and 1,566 Test_llama inputs. The MAE change remains positive for all four editing extent scorers, with paired 95% intervals above zero. The bootstrap uses 3,000 resamples of human source groups and keeps the clean and adversarial measurements together.

Ablation Study We compare three variants on Test. *No prior* removes the paired humanization prior, *Prior only* removes the surrogate model ranking, and *Full method* uses both signals. As shown in Table 15, on Llama, the full method achieves the largest mean reduction and the highest values for all three reduction rates. On RoBERTa, the full method gives the highest proportion of score decreases. These results support the complementary roles of the paired humanization prior and surrogate model.

D Qualitative Examples

The following examples illustrate the changes made by PRESS. Orange marks replaced wording in the input and teal marks new wording in the output. All scores and edit rates refer to complete texts. The first example shows the full text. The remaining examples show the modified sentences, with

Table 15: Ablation results on Test.

	Llama				RoBERTa			
Configuration	Mean Δ_T	$P_{\downarrow}$	$P_{10\%}$	$P_{20\%}$	Mean Δ_T	$P_{\downarrow}$	$P_{10\%}$	$P_{20\%}$
No prior	0.06789	73.9%	38.9%	23.4%	0.01613	70.5%	15.1%	8.2%
Prior only	0.05927	73.6%	36.6%	21.2%	0.02131	73.4%	17.2%	9.8%
Full method	**0.07235**	**75.7%**	**40.9%**	**24.6%**	**0.02047**	**74.3%**	16.6%	9.5%

omitted unchanged context indicated by · · · . Local F1 denotes the mean of the recorded sentence similarity checks.

D.1 A VISITOR REVIEW

Input

This is a must-see. The tour guide was knowledgeable, and the I-max type theater showing Hurricane was outstanding. On site, **you'll** find two museums, a theater, and a water park. The Spruce Goose is breathtaking **- massive** and open for interior viewing. With mostly handicapped-accessible buildings, you can easily explore the vast collection of planes, from vintage to modern, on display **everywhere.** The uniforms and historical progression are fascinating. Don't miss it!

Attack output

This is a must-see. The tour guide was knowledgeable, and the I-max type theater showing Hurricane was outstanding. On site, **visitors will** find two museums, a theater, and a water park. The Spruce Goose is breathtaking and **massive,** open for interior viewing. With mostly handicapped-accessible buildings, you can easily explore the vast collection of planes, from vintage to modern, on display **all around.** The uniforms and **the** historical progression are fascinating. Don't miss it!

	EditLens Llama	EditLens RoBERTa	D_{CD}	Edit rate
Input	0.259	0.446	0.0811	—
Output	0.206	0.298	0.0847	11.11%

Local F1: 0.966. Input length: 72 words.

D.2 AN EXPLANATION OF BIRD EVOLUTION

Input

Scientists believe birds first appeared **roughly** 150 million years ago, evolving from small feathered carnivorous dinosaurs.

. . .

These rapid changes **have** made bird evolution **hard** to **pin down.**

. . .

But paradoxes **abound** within these groupings.

. . .

Reddy **acknowledges** the results will **likely stir** debate, but she remains confident in the findings.

Attack output

Scientists believe birds first appeared **about** 150 million years ago, evolving from small feathered carnivorous dinosaurs.

. . .

These rapid changes made bird evolution **difficult** to **determine.**

. . .

But **there are** paradoxes within these groupings.

. . .

Reddy **says** the results **probably** will **cause** debate, but she remains confident in the findings.

	EditLens Llama	EditLens RoBERTa	D_{CD}	Edit rate
Input	0.352	0.304	0.0203	—
Output	0.290	0.226	0.0194	3.81%

Local F1: 0.919. Input length: 341 words.

D.3 A NEWS REPORT

Input

A British band and their manager died in a car **accident** in Sweden on **13 February.**

. . .

The police **said** the driver had no alcohol or drugs in their system.

Attack output

A British band and their manager died in a car **crash** in Sweden on **February 13th.**

. . .

The police **stated** the driver had no alcohol or drugs in their system.

	EditLens Llama	EditLens RoBERTa	D_{CD}	Edit rate
Input	0.627	0.538	0.2275	—
Output	0.551	0.420	0.2392	3.81%

Local F1: 0.975. Input length: 105 words.

D.4 AN AI GENERATED NARRATIVE

Input

As I stood frozen in my doorway, the little man **handed** me a glowing envelope with an unsettling smile.

. . .

His eyes, which seemed to **be taking** up half of his face, sparkled with an otherworldly intensity.

. . .

The humming grew louder, and the air around me **began** to distort and **ripple.**

. . .

When I opened my eyes, I **found** myself in a vast, empty space filled with rows of identical, glowing envelopes.

Attack output

As I stood frozen in my doorway, the little man **gave** me a glowing envelope with an unsettling smile.

. . .

His eyes, which seemed to **take** up half of his face, sparkled with an otherworldly intensity.

. . .

The humming grew louder, and the air around me **started** to distort and **ripple noticeably.**

. . .

When I opened my eyes, I **discovered** myself in a vast, empty space filled with rows of identical, glowing envelopes.

	EditLens Llama	EditLens RoBERTa	D_{CD}	Edit rate
Input	0.775	0.662	0.3004	—
Output	0.664	0.645	0.2966	1.61%

Local F1: 0.981. Input length: 435 words.

AI USE STATEMENT

AI tools are used only to aid and polish the writing.