Text Length and the Reliability of Human versus AI Text Classification in Academic Writing

TextPulse Research. Working Paper.

Abstract

Our previous study classified human-written and AI-rewritten academic texts from interpretable stylometric features with an AUC of 0.936 on texts with a median length near 290 words. This study measures how that reliability depends on text length. From 25,561 text pairs in which both the human text and its AI rewrite contain at least 300 words, we cut every text to 25, 50, 100, 150, 200, 250, and 300 words, re-extract 45 stylometric features at each length, and retrain the classifier at each length under grouped five-fold cross-validation. The composition of the sample is identical at every length, so the results isolate text length from every other property of the texts. Classification reliability increases monotonically with length. The AUC of a logistic regression increases from 0.817 at 25 words to 0.940 at 300 words and 0.950 on the untruncated full texts, and most of the increase is at the 150-word length. Reliable individual verdicts degrade faster than the AUC suggests. Identifying 95 percent of AI rewrites requires falsely flagging 34 percent of human texts at 300 words, but 68 percent at 25 words, and a strict one percent false positive budget identifies only 17 percent of AI rewrites at 25 words. A classifier trained on full-length texts and applied to short texts at a fixed threshold flags 70 percent of 25-word human texts as AI while its detection of AI rewrites hardly changes, since short texts of any authorship have mechanically high lexical diversity, which the classifier reads as an AI characteristic. Word-level features preserve most of their distinguishing power at short lengths, and sentence-level variation carries almost no signal under 50 words, where texts average fewer than three sentences. All per-text scores and analysis code are made publicly available.

1. Introduction

Our prior study treated the separation of human-written and AI-rewritten academic text as a supervised classification task on interpretable stylometric features and reported an AUC of 0.936 with a documented error structure [1]. Those numbers describe texts whose median length is near 290 words. In practice, the texts submitted to AI detection tools are often far shorter, namely, single paragraphs, abstracts, discussion-board posts, or isolated sections of a longer document. Commercial AI detection tools acknowledge the problem in general terms. GPTZero states that the accuracy of its model increases as more text is submitted, and that document-level classification is more accurate than paragraph-level, which is in turn more accurate than sentence-level classification [2]. Detection theory makes the same prediction, since the distinguishability of two text distributions increases with the number of observed tokens [3, 4].

Published accuracy figures for AI text detection, including ours, are tied to the length distribution of the corpus they were computed on, and comparisons between studies mix text length with topic, genre, model, and method. This study removes this mix with a truncation design. We keep the texts fixed and vary only how much of each text the classifier is allowed to see at once. Every result at 25 words and every result at 300 words describes the same 25,561 text pairs, the same models, the same disciplines, and the same classification protocol.

The study answers four questions. First, how does classification reliability increase with text length when the classifier is trained and evaluated at the same length? Second, how do the operating points that matter in practice, namely, high recall of AI text or strict false positives, degrade as texts become shorter? Third, what happens in the realistic deployment case where a classifier trained on full-length documents is applied to short texts without recalibration? Fourth, which stylometric markers survive shortening and which become unusable?

The fourth question connects to the feature findings of our earlier studies, which measured what AI rewriting does to vocabulary [5], to sentence-length burstiness [6], and to the broader stylometric profile [7]. A short text contains a few sentences, so any feature defined on variation between sentences (e.g., burstiness) loses its effect, while rates defined on individual words remain measurable.

2. Related work

The dependence of detection on text length has been noted across the literature, but rarely isolated. Ippolito et al. reported that both human raters and automatic classifiers identify generated text more accurately as excerpt length increases [8]. The report accompanying the staged release of GPT-2 quantified the effect for a fine-tuned RoBERTa detector, whose accuracy becomes higher for longer text and passes 90 percent at roughly 70 words [9]. Chakraborty et al. gave the observation a theoretical basis, and showed that the detection of AI text remains possible whenever the human and machine distributions differ, with the required amount of text growing as the distributions approach each other [4]. Sadasivan et al. derived the corresponding limit, in which the AUROC of the best possible detector is bounded by the total variation distance between the human and machine text distributions [10]. The survey by Fraser et al. gathers the empirical thresholds reported so far, namely, on the order of 100 words for statistical and fine-tuned classifiers to work accurately, roughly 120 words for several standard detectors to reach their full performance, and about 200 words for detecting the strongest models, while watermarking schemes need only tens of words [3]. The watermark of Kirchenbauer et al. is the main example, since it embeds a signal designed to be algorithmically detectable from a short span of tokens [11]. Our results below are consistent with those thresholds and add the error anatomy at each predefined length.

Short texts have been studied most directly on social media. Kumarage et al. showed that stylometric features improve the detection of AI-generated tweets, where standard language-model detectors struggle with the short input [12]. Przystalski et al. classified ten-sentence samples of

human and AI text with tree-based models on stylometric features and reported binary accuracies between 0.79 and 1.0 depending on the model pair [13]. Both studies support the finding that surface style carries usable signal in short texts. Neither holds the texts constant while varying length, which is the key contribution of this study. A related line of work moves detection below the document level entirely. Wang et al. introduced a sentence-level detection task, found that methods built for whole documents struggle on single sentences, and trained a sequence-labeling detector on token probability features to handle them [14].

Evaluations of commercial detectors offer the practical context. Weber-Wulff et al. found no tool exceeding 80 percent accuracy on a mixed corpus [15], and Liang et al. showed that detector false positives concentrate on non-native English writers [16]. The classifier OpenAI released in 2023 made the length constraint explicit. It required at least 1,000 characters of input, was described by OpenAI itself as very unreliable on short texts, with a reliability that typically improves as the input text lengthens, and was withdrawn within six months for its low rate of accuracy [17]. Our prior study documented the same concentration of false positives on formal human writing for an open stylometric classifier [1]. This study shows that text length is a second, independent axis along which false positives concentrate, and one that operates mechanically rather than through writing style.

3. Data and methods

3.1 Corpus and cohort

The data are the human-AI rewrite pairs of our prior studies [5, 6, 7], namely 60,786 pairs of human-written academic text and an AI rewrite of that text, produced by eight AI model configurations across six model families in two corpora. Human texts have a median length of 295 words and AI rewrites of 284 words.

Truncating to a target length requires texts at least that long, and selecting longer texts at longer bands would change the composition of the sample across bands. We therefore fix a cohort, namely, the 25,561 pairs in which both the human text and the AI rewrite contain at least 300 words. Every length band evaluates exactly this cohort. After deduplication of repeated human sources, the cohort classification table holds 50,701 texts, of which 25,140 are distinct human texts and 25,561 are AI rewrites. As a robustness check we repeat the analysis on all pairs eligible at each band, which uses up to 121,098 texts at the shortest bands.

3.2 Truncation

Each text is cut at the end of its 25th, 50th, 100th, 150th, 200th, 250th, and 300th word, and the feature extractor of the prior study is rerun on each cut [7]. The cut falls at a word boundary, so the final sentence of a truncated text is usually incomplete, which matches how excerpts are pasted

into detection tools in practice. A text cut to 25 words averages 1.6 sentences, at 50 words 2.7 sentences, and at 300 words 13.4 sentences. The untruncated texts are retained as an eighth band.

3.3 Features

We use 45 of the 49 stylometric features defined in the prior study [7]. The em dash and en dash rates are excluded, as in the prior studies [1, 7]. The raw word and sentence counts are additionally excluded, because truncation sets them by construction and a classifier that reads them would be measuring the band rather than the writing. Our prior study showed that removing these two features changes the full-corpus AUC only from 0.936 to 0.934, so the exclusion costs almost nothing [1]. As a validation, our protocol on the untruncated full corpus with the 45 features reproduces that number exactly, with an AUC of 0.934 and an accuracy of 86.2 percent on 121,092 texts.

3.4 Classification protocol

The primary model is a logistic regression on standardized features, with a histogram gradient boosting classifier as a stronger non-linear reference, both with balanced class weights (scikit-learn 1.9.0). All results come from five-fold cross-validation grouped by human source text, so no source or any rewrite of it appears in both training and evaluation, and scores are pooled out of fold. The folds contain the same texts in every band.

Three designs are evaluated. In the matched design, the classifier is trained and evaluated on texts of the same band, which measures the reliability attainable when the training data matches the deployment length. In the mismatch design, the classifier is trained on the untruncated texts of the training folds and evaluated on the truncated texts of the evaluation folds, which measures what happens when a detector built on long documents receives short ones. In the eligible design, the matched analysis is repeated on all pairs long enough for each band regardless of cohort membership.

4. Reliability as a function of length

4.1 Overall separability

Table 1 presents the matched results. The AUC of the logistic regression increases monotonically from 0.817 at 25 words to 0.940 at 300 words, and the gradient boosting reference increases from 0.821 to 0.957. The increase is steepest below 100 words. Between 25 and 100 words the logistic regression gains 8.5 points of AUC, between 100 and 200 words 2.6 points, and between 200 and 300 words 1.3 points. On the untruncated cohort texts, which average 466 words on the human side, the models reach 0.950 and 0.968.

Length (words)	LR AUC	LR accuracy	GB AUC	GB accuracy
25	0.817	73.9%	0.821	74.1%
50	0.863	78.2%	0.871	78.7%
100	0.902	82.3%	0.913	83.1%
150	0.918	84.0%	0.932	85.7%
200	0.928	85.2%	0.943	87.0%
250	0.935	86.1%	0.952	88.3%
300	0.940	86.8%	0.957	89.1%
Untruncated	0.950	88.4%	0.968	90.9%

Figure 1 shows the score distributions at 50 and 300 words. At 300 words the two classes concentrate at their correct ends of the scale with a thin overlap region. At 50 words both distributions flatten and spread toward the middle, and the overlap region holds a large share of both classes. The information that pushes a text confidently to one end of the scale accumulates with length, and 50 words do not supply much of it.

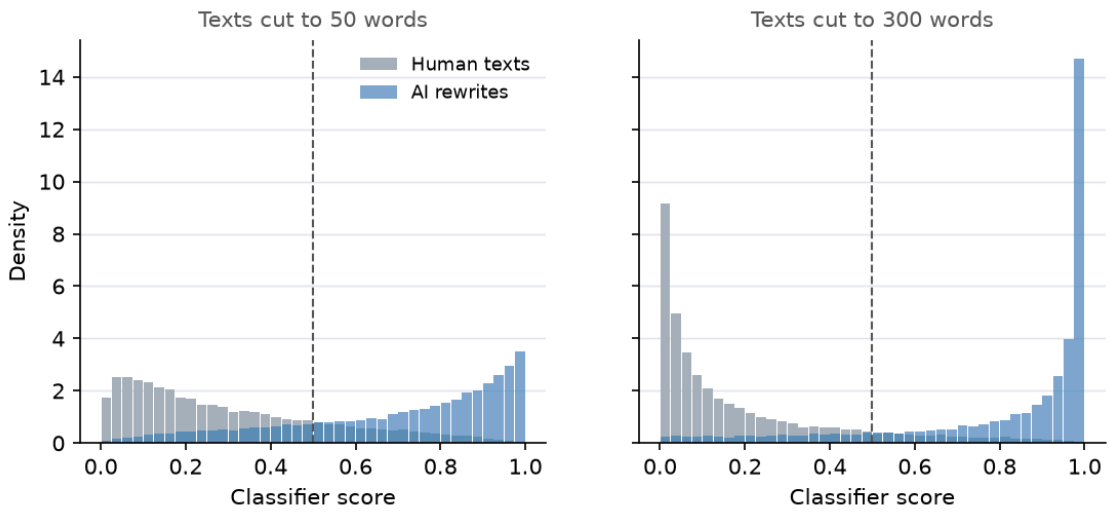

Figure 1. Classifier score distributions for human texts and AI rewrites at 50 and 300 words, pooled out-of-fold scores, matched design.

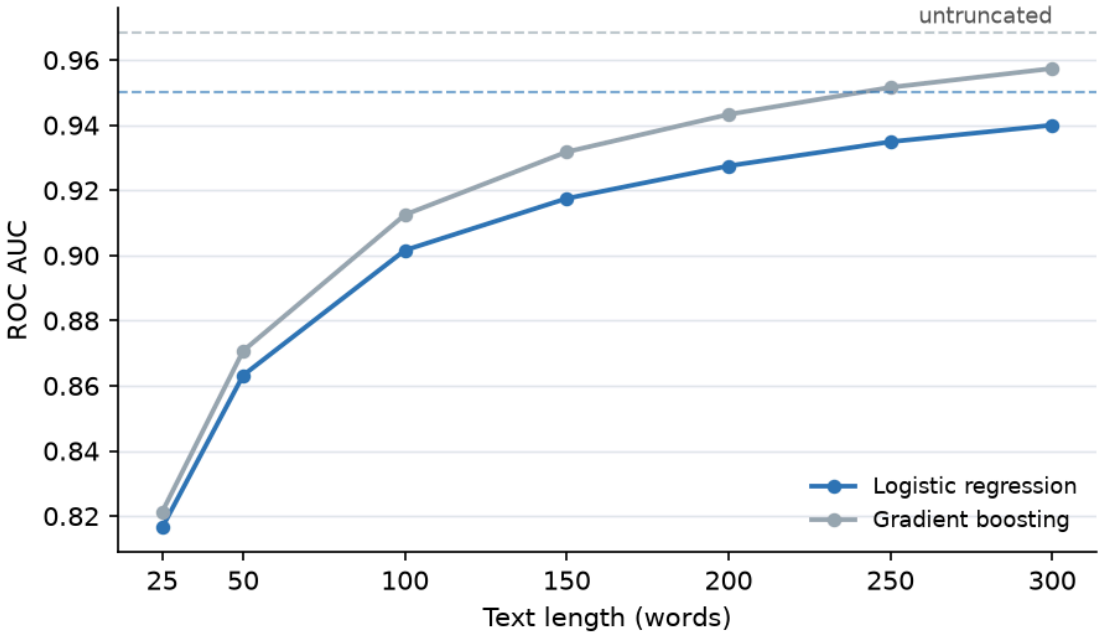

Figure 2. ROC AUC against text length for the logistic regression and the gradient boosting reference, matched design. Dashed lines mark the untruncated cohort.

4.2 Operating points degrade faster than the AUC

The AUC understates the practical loss, since usable verdicts require scores near the ends of the scale, and shortening moves both classes toward the middle. Table 2 presents the operating points of the logistic regression at each length.

Length (words)	Humans flagged at 90% recall	Humans flagged at 95% recall	AI identified at 1% false positives
25	51.7%	67.6%	17.2%
50	42.6%	58.8%	28.0%
100	32.5%	49.4%	40.5%
150	27.0%	44.1%	47.9%
200	23.0%	40.1%	50.5%
250	20.5%	36.7%	54.2%
300	18.8%	33.9%	56.5%
Untruncated	14.6%	28.6%	60.6%

Reading the table from the bottom up, identifying 95 percent of AI rewrites requires falsely flagging 29 percent of human texts on full-length documents, 34 percent at 300 words, half of them at 100 words, and two thirds at 25 words. Conversely, capping false flags at one percent of human texts

identifies 61 percent of AI rewrites on full-length documents but 17 percent at 25 words. At paragraph length, a strict false positive budget makes the classifier miss most AI text, and a high recall requirement makes it flag most human text. Figure 3 shows these curves.

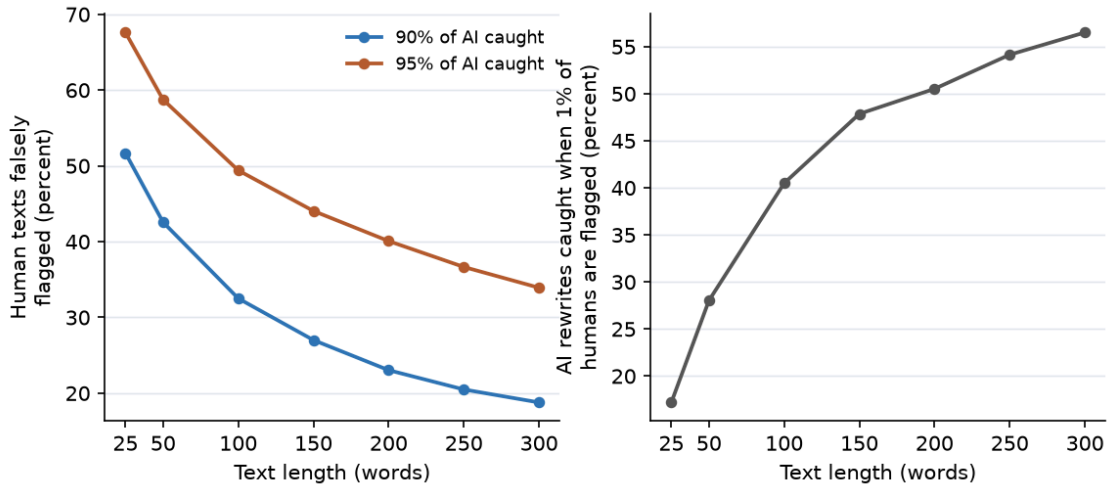

Figure 3. Operating points against text length, matched design. Left, the share of human texts falsely flagged when the threshold is set to identify 90 and 95 percent of AI rewrites. Right, the share of AI rewrites identified when the threshold is set to flag one percent of human texts.

4.3 Robustness

The eligible design, which uses every pair long enough for each band rather than the fixed cohort, offers the same concept, with AUCs from 0.793 at 25 words to 0.940 at 300 words. The values at short bands are slightly lower than in the cohort since the eligible sample adds the shorter, more heterogeneous texts. The cohort itself is also informative about composition. Its untruncated AUC of 0.950 exceeds the full-corpus 0.934, so even without any truncation, longer academic texts are easier to classify than shorter ones.

5. The deployment mismatch

A deployed AI detector is rarely retrained for each input length. It is trained once, mostly on long documents, and then receives whatever users paste into it. The mismatch design measures this case directly, with the same folds and texts as the matched design.

5.1 Ranking quality

Figure 4 compares the two designs. The mismatch AUC is lower at every length, and the gap widens as texts shorten, from 0.7 points of AUC at 300 words to 7.3 points at 25 words, where the matched classifier reaches 0.817 and the mismatched one 0.743. Training on long texts therefore fails to

prepare the classifier for short ones, and it actively costs ranking quality where reliability is already lowest.

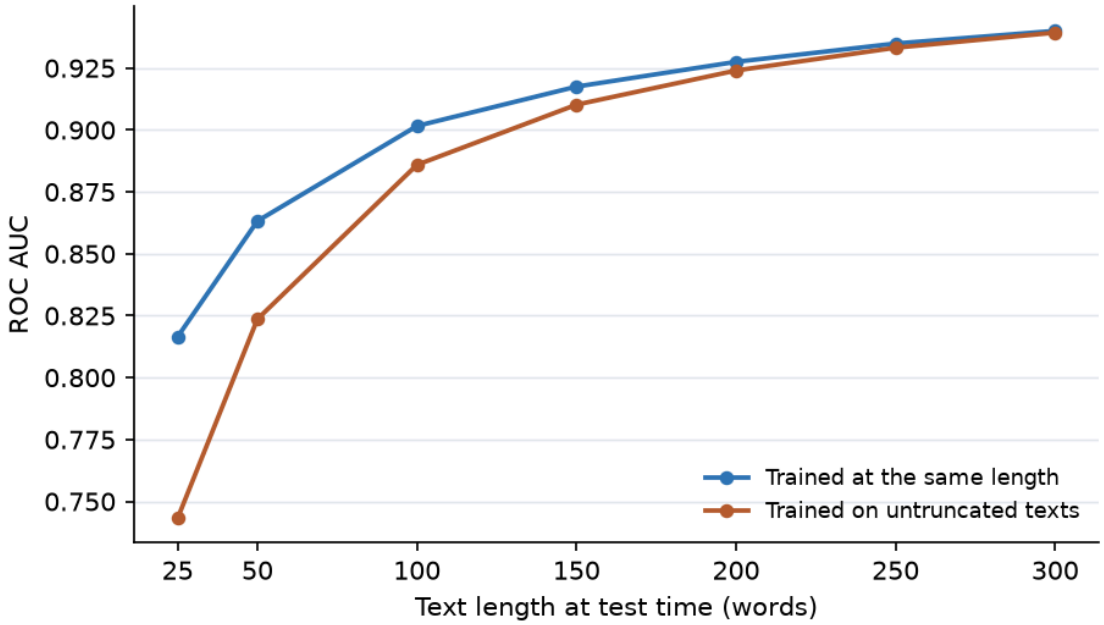

Figure 4. ROC AUC against text length at evaluation time, for a classifier trained at the same length and a classifier trained on untruncated texts.

5.2 The errors drift onto human writers

The more consequential effect appears at a fixed decision threshold. Figure 5 shows the two error rates of the full-length-trained classifier at the default threshold as a function of input length. Detection of AI rewrites barely moves, from 88.4 percent on 300-word inputs to 92.4 percent on 25-word inputs. The false positive rate on human texts increases more than four times, from 16.1 percent at 300 words to 39.1 percent at 100 words and 69.8 percent at 25 words. Applied without recalibration, the classifier does not become noisy on short texts. It becomes systematically biased toward calling them AI.

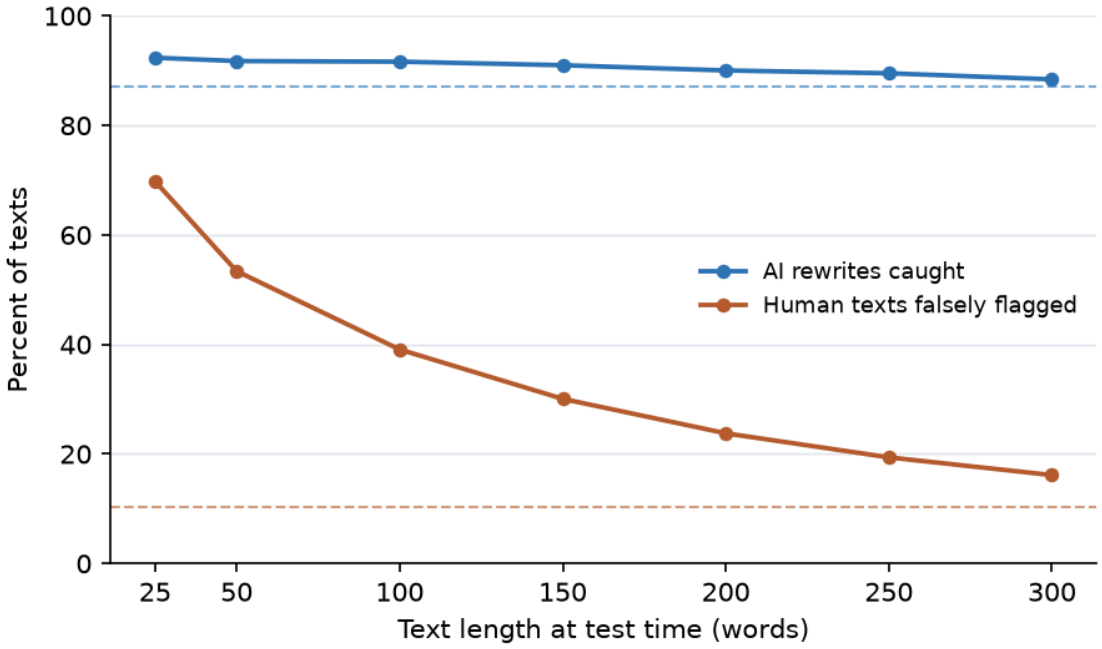

Figure 5. Error rates of a classifier trained on untruncated texts and applied at the default threshold to shorter inputs. Dashed lines mark the same classifier on untruncated inputs.

The mechanism is mechanical rather than stylistic. Type-token ratio, one of the strongest AI markers in this feature set, increases as texts get shorter for both classes, simply because a short sample offers fewer chances to repeat a word. In the cohort, human texts average a type-token ratio of 0.47 untruncated and 0.87 at 25 words, while full-length AI rewrites average 0.55. A classifier calibrated on full-length texts learned that a ratio near 0.55 separates AI from human. Every 25-word input, whatever its authorship, has a ratio far above that mark and is pushed toward the AI side of the scale. The same logic applies to the moving-average type-token ratio, whose 50-word window exceeds the text itself at the shortest bands.

This result generalizes the false positive anatomy of our prior study, which found that misclassified human texts are the most formal ones [1]. Formality is a property of the writer. Length is a property of the submission. A human writer with an unremarkable style can still be flagged reliably by a length-mismatched detector simply by submitting one paragraph instead of multiple.

6. Which markers survive shortening

Figure 6 tracks the separation of the strongest features across lengths, measured as the absolute standardized mean difference between classes. The features fall into three groups.

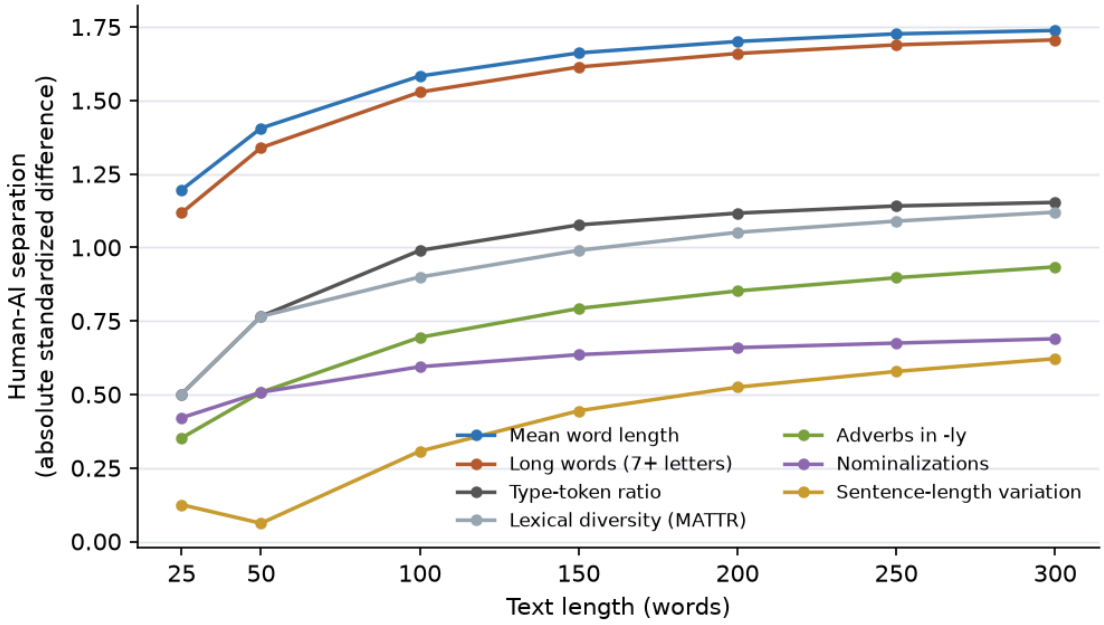

Figure 6. Human-AI separation of the top features against text length, cohort, absolute standardized mean difference.

Word-level rates are robust. Mean word length and the rate of long words separate the classes at 1.75 and 1.72 standard deviations on untruncated texts and still at 1.20 and 1.12 at 25 words, retaining about two thirds of their power on a single sentence or two. The -ly adverb and nominalization rates decay more but remain usable.

Diversity measures are length-dependent but not destroyed. Type-token ratio separates at 1.19 untruncated and 0.50 at 25 words. The measure remains informative within a band, where all texts share one length, even though its level moves with length, which is what breaks the mismatched classifier of Section 5.

Sentence-level variation disappears. The coefficient of variation of sentence length, the burstiness property our prior study found compressed by every AI configuration [6], separates the classes at 0.71 standard deviations on untruncated texts, 0.31 at 100 words, and effectively zero at 50 words, below which its sign flips as the statistic degenerates on samples of one or two sentences. Burstiness is effective, but it is not measurable in a short paragraph. Any detector relying on sentence-length variation needs several sentences to reliably measure it.

Per-model detectability compresses in the same manner. At 300 words the miss rate at the default threshold ranges from 5.2 percent for DeepSeek rewrites to 34.9 percent for Qwen, an ordering that tracks how much each model changes its input [7]. At 100 words every model except Qwen becomes harder to catch, and the ordering compresses, with DeepSeek chat reaching 35.6 percent. Figure 7 shows all eight configurations. The models easiest to detect at length lose the most detectability when texts shorten, since much of their signature is in the sentence-level and diversity features that shortening removes.

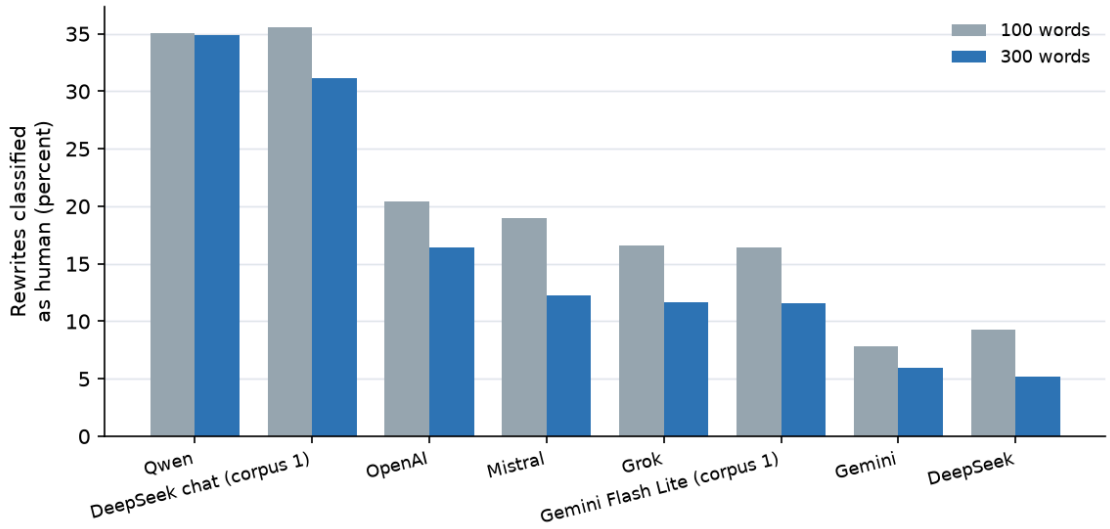

Figure 7. The share of each configuration's rewrites classified as human at the default threshold, at 100 and 300 words, matched design.

7. Discussion

Several key points are worth noting. First, minimum length requirements are justified, and our numbers state where the reliability is lost. Below roughly 100 words, classification from surface style loses reliability quickly, and below 50 words the usable operating points are rendered useless, in line with the thresholds collected across the literature [3] and with vendor guidance [2]. An abstained verdict on a relatively short text is more reliable than a scored one.

Second, a score on a short text is not a weaker version of a score on a long text. It is a different measurement with a different error structure. In the mismatch design the false positive rate at a fixed threshold quadruples between 300 and 25 words while recall of AI text is unchanged. A detector that does not recalibrate by length considers short human submissions as AI written. Combined with the base rate arithmetic of our prior study, in which most flagged texts can be human even at full length when the AI share of the population is low [1], short inputs move the reliability of an individual flag toward zero.

Third, the asymmetry matters for fairness. The false positives of our prior study concentrated on formal writers [1], and those of Liang et al. on non-native writers [16]. Length-mismatched deployment adds a procedural bias against anyone evaluated on a paragraph, independent of who they are or their writing style. Unlike style, submission length is often set by the evaluator rather than the writer, which makes this failure mode more arbitrary and easier to fix. Requiring more text or refusing a verdict below a stated length removes it entirely.

8. Limitations and conclusion

Truncation approximates short documents with cut-off long ones. Natively short writing, such as an answer written to fit a word limit, may differ stylistically from the opening of a longer document, so our short-band accuracies should be read as measurements of the length effect in isolation rather than of any specific short-text population. The cohort holds composition fixed at the price of describing the longer texts of the corpus, though the eligible analysis reproduces every trend on the full data. The corpus is AI rewriting of English academic text, and the eight configurations were captured at fixed points in 2026. The classifier is a research instrument, and nothing here implies that commercial AI detectors use these features or thresholds.

Text length is a first-order determinant of AI text detectability from surface style. The same texts, AI models, and protocol yield an AUC from 0.82 to 0.95, depending only on how much of each text the classifier sees. Most of the reliability is at roughly 150 words. The practically usable operating points require at least roughly 100, and applying a detector trained on full-length texts to short texts converts the length deficit into false positives (of human-written texts seen as AI-generated) at up to 70 percent. Any accuracy claim for AI text detection that does not state the text length it was measured at is incomplete.

Data availability

Per-text classifier scores for all cohort texts at every length band (opaque pair identifiers, corpus, model, discipline, side, and the out-of-fold scores of the matched and mismatched models), the per-band metrics, and all analysis and figure code are openly available on Zenodo at https://doi.org/10.5281/zenodo.22050346.

References

[1] TextPulse Research (2026). Human versus AI text classification from stylometric features across 121,092 academic texts. Working paper. https://textpulse.ai/research/human-vs-ai-text-classification. https://doi.org/10.5281/zenodo.22048476

[2] GPTZero (2026). What are the limitations of GPTZero's AI classifier? Support article, accessed August 22, 2026. https://support.gptzero.me/hc/en-us/articles/15129396117143-What-are-the-limitations-of-GPTZero-s-AI-classifier

[3] Fraser, K. C., Dawkins, H., and Kiritchenko, S. (2025). Detecting AI-generated text: Factors influencing detectability with current methods. Journal of Artificial Intelligence Research, 82, 2233-2278. https://arxiv.org/abs/2406.15583

[4] Chakraborty, S., Bedi, A. S., Zhu, S., An, B., Manocha, D., and Huang, F. (2024). Position: On the possibilities of AI-generated text detection. Proceedings of the 41st International Conference on Machine Learning (ICML 2024), PMLR 235, 6093-6115. https://proceedings.mlr.press/v235/chakraborty24a.html

[5] TextPulse Research (2026). The vocabulary fingerprint of AI rewriting: common words AI language models prioritize. Working paper. https://textpulse.ai/research/ai-vocabulary-fingerprint. https://doi.org/10.5281/zenodo.22028377

[6] TextPulse Research (2026). Sentence-length burstiness as a cross-disciplinary and cross-model signal of AI rewriting. Working paper. https://textpulse.ai/research/ai-burstiness-sentence-length. https://doi.org/10.5281/zenodo.22033062

[7] TextPulse Research (2026). Stylometric fingerprints of AI rewriting: punctuation, syntax, and model attribution across 60,786 paired texts. Working paper. https://textpulse.ai/research/ai-style-fingerprints. https://doi.org/10.5281/zenodo.22040683

[8] Ippolito, D., Duckworth, D., Callison-Burch, C., and Eck, D. (2020). Automatic detection of generated text is easiest when humans are fooled. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), 1808-1822. https://aclanthology.org/2020.acl-main.164/

[9] Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., and Wang, J. (2019). Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203. https://arxiv.org/abs/1908.09203

[10] Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. (2025). Can AI-generated text be reliably detected? Stress testing AI text detectors under various attacks. Transactions on Machine Learning Research. https://arxiv.org/abs/2303.11156

[11] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. (2023). A watermark for large language models. Proceedings of the 40th International Conference on Machine Learning (ICML 2023), PMLR 202, 17061-17084. https://proceedings.mlr.press/v202/kirchenbauer23a.html

[12] Kumarage, T., Garland, J., Bhattacharjee, A., Trapeznikov, K., Ruston, S., and Liu, H. (2023). Stylometric detection of AI-generated text in Twitter timelines. arXiv preprint arXiv:2303.03697. https://arxiv.org/abs/2303.03697

[13] Przystalski, K., Argasiński, J. K., Grabska-Gradzińska, I., and Ochab, J. K. (2026). Stylometry recognizes human and LLM-generated texts in short samples. Expert Systems with Applications, 296, 129001. https://doi.org/10.1016/j.eswa.2025.129001

[14] Wang, P., Li, L., Ren, K., Jiang, B., Zhang, D., and Qiu, X. (2023). SeqXGPT: Sentence-level AI-generated text detection. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023), 1144-1156. https://aclanthology.org/2023.emnlp-main.73/

[15] Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., and Waddington, L. (2023). Testing of detection tools for AI-generated text.

International Journal for Educational Integrity, 19, 26. https://doi.org/10.1007/s40979-023-00146-z

[16] Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

[17] OpenAI (2023). New AI classifier for indicating AI-written text. Published January 31, 2023, updated July 20, 2023, accessed August 22, 2026. https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/