The Detectability of Partially AI-Rewritten Academic Documents from Stylometric Features

TextPulse Research. Working Paper.

Abstract

Real documents are often partly AI-processed, with a few sections AI-rewritten and the rest left as human-written. This study measures how a stylometric human-versus-AI binary responds as the AI-rewritten share of a document increases from 0 to 100 percent. From 25,561 content-aligned pairs of human-written academic texts and their AI rewrites, we splice documents in which a controlled fraction of the words, in increments of 10 percent, has been rewritten by one of eight AI model configurations, and score all 382,993 resulting documents with a classifier trained only on fully human and fully AI-rewritten texts, under grouped cross-validation throughout. The classifier reads the AI share as a proportion on a continuum, rather than detecting its presence and categorizing it into a discrete class of AI or Human. The mean score increases almost linearly from 0.17 to 0.83, and the share of documents flagged at the default threshold crosses 50 percent almost exactly at half AI content. Scattering the AI sentences through the document instead of concentrating them in one block leaves the response unchanged. Small AI shares are close to invisible. A document with 10 percent of its words AI-rewritten is separated from fully human documents with an AUC of 0.57, barely above chance, and under a strict one percent false positive budget only 15 percent of half-rewritten documents are flagged. Every feature mixes near linearly with the AI share except sentence-length variation, which is above the value linear mixing predicts, since the style change at the splice point itself adds variation, and 36 percent of half-rewritten documents show more sentence-length variation than the original human text. Distinguishing half-rewritten documents from pure ones as a class reaches an AUC of only 0.59 for a linear model and 0.76 for gradient boosting, so partial rewriting is detectable mainly as a lower score, not as a recognizable discrete category. Per-document scores and all analysis code are made publicly available.

1. Introduction

Our prior studies classified fully human-written against fully AI-rewritten academic texts from interpretable stylometric features, reaching an AUC of 0.936 across 121,092 texts [1] and characterizing how that number depends on text length [2]. Both studies share an assumption with most of the detection literature, namely, that a document belongs entirely to one discrete class. Realistic usage, however, is not always a full AI generation or rewrite. A writer drafts a document, passes some paragraphs through an AI rewriting model (ChatGPT, Claude, Gemini, etc.) or paraphrasing tool (ProofreaderPro, Grammarly, QuillBot, etc.), and submits the result. The

document is then neither human nor AI, and a detector that scores it must respond with sensitivity to an AI-human mixture.

Our detector-agreement study touched this case briefly, with 30 hybrid texts built from half human and half AI-generated content, and found that nine commercial detectors treated them inconsistently, with verdicts scattered across the full range [3]. A controlled measurement of the mixture response, namely, how the score of a fixed classifier moves as the AI share of a document increases in predefined steps, with everything else held constant, merits further investigation.

The corpus of our prior studies contains 25,561 pairs in which a human academic text of at least 300 words was rewritten in full by an AI model, which makes the two versions of every document content-aligned. Splicing the human version up to a chosen position onto the AI version from that position onward yields a document in which a known fraction of the words has been AI-rewritten while the content stays continuous. We build such documents at every AI share from 10 to 90 percent in steps of 10, score them with the classifier of the prior studies trained on pure texts only, and measure the response.

The result is a response curve on a spectrum, not a detection event. The classifier output tracks the AI share almost linearly along its entire range. This has direct consequences for how document-level AI scores should be read, and the experiments below quantify them, including the share below which partial rewriting is effectively undetectable, the behavior of practical thresholds on mixed documents, the one feature that does not mix linearly, and the difference between AI models.

2. Related work

The mixed-document case has been studied mainly as a boundary-detection task. Dugan et al. measured whether human readers can locate the point where a document transitions from human-written to machine-generated, and found the task difficult, with wide variance between annotators [4]. SemEval-2024 Task 8 posed the same problem to machines, asking systems to identify the word position where authorship changes within a mixed text [5]. Kumarage et al. detected the point in a social media timeline where AI-generated posts begin, using stylometric features [6]. These works attempt to detect where the transition occurs. This study asks the prior question, namely, what a whole-document score even means when the document is mixed, since the whole-document score is what AI detection tools return.

The construction we use follows the hybrid concept of our detector-agreement study, in which 30 documents splicing human and AI halves received intermediate and inconsistent verdicts from nine commercial detectors [3]. The present design scales that probe from 30 hand-built documents at one mixing ratio to 306,732 spliced documents across nine ratios, with a classifier whose training data, features, and thresholds are fully known.

The feature set and classification protocol come from our prior studies [1, 7], the length dependence of the pure-class problem is quantified in [2], and the sentence-length variation

feature that behaves distinctively here was established as a cross-model marker of AI rewriting in [8].

3. Data and methods

3.1 Corpus and document construction

The data are the 25,561 pairs of our length study cohort [2], namely, every pair in the corpus of the prior studies in which both the human text and its AI rewrite contain at least 300 words. The rewrites come from eight AI model configurations across six model families in two corpora [1, 7].

For each pair and each target share f, the mixed document takes the human text up to the sentence boundary nearest (1 - f) of the human text's words, followed by the AI rewrite from the sentence boundary nearest (1 - f) of the rewrite's words. Since the AI rewrite is a full rewrite of the same source, content stays continuous across the splice, and the last f of the document's words are AI-rewritten, up to sentence-boundary rounding. The achieved AI share tracks the target closely, averaging 9.4 percent at the 10 percent target and 89.3 percent at the 90 percent target. We build suffix documents at f from 10 to 90 percent in steps of 10, keep both pure versions of every pair, and additionally build prefix documents at f of 25, 50, and 75 percent, in which the AI part comes first, as a position check. We also build one interleaved document per pair at approximately half AI content, in which every second sentence of the human text is replaced by the sentence of the AI rewrite at the matching position, described in Section 7. In total 382,993 documents are scored.

3.2 Features and protocol

Each document is described by the 45 stylometric features of the length study, namely, the feature set of the prior studies without the two dash rates and without the raw word and sentence counts [1, 2, 7]. The classifier is the same logistic regression on standardized features, with a histogram gradient boosting model as the non-linear reference (scikit-learn 1.9.0, balanced class weights).

Training uses pure texts only, namely, the deduplicated human texts and the full rewrites. All models are trained under five-fold cross-validation grouped by human source, and every document, pure or mixed, is scored by the fold model that never saw its source text or any rewrite of it. As validation, the pure ends reproduce the untruncated cohort numbers of the length study exactly, with an AUC of 0.950 for the logistic regression and 0.968 for the gradient boosting model [2].

3.3 Thresholds

Three thresholds are evaluated, all of which are fixed from the pure out-of-fold scores before any mixed document is examined. The default threshold of 0.5 flags 10.3 percent of pure human texts and 87.0 percent of pure rewrites. The threshold set to identify 95 percent of pure rewrites flags

28.6 percent of pure human texts. The threshold set to flag 1 percent of pure human texts identifies 60.6 percent of pure rewrites.

4. The response to partial rewriting

4.1 The score reads as a proportion

Table 1 and Figure 2 present the main results. The mean classifier score increases from 0.172 on fully human documents to 0.828 on fully AI-rewritten ones, and it does this almost linearly, gaining close to 0.066 for every additional 10 points of AI share. The share of documents flagged at the default threshold crosses 50 percent at an AI share of 50.4 percent. There is no share below which the classifier is blind followed by a share where it reacts. Every additional block of rewritten text moves the score by about the same amount.

AI share (percent)	Mean score	Flagged, default	Flagged, 95% recall threshold	Flagged, 1% FPR threshold	AUC against fully human
0 (fully human)	0.172	10.3%	28.6%	1.0%	
10	0.222	15.5%	37.6%	1.8%	0.565
20	0.280	22.0%	46.6%	3.2%	0.630
30	0.347	29.9%	56.3%	5.5%	0.691
40	0.421	39.9%	65.3%	9.2%	0.747
50	0.498	49.6%	73.4%	14.9%	0.798
60	0.577	59.5%	80.3%	22.5%	0.842
70	0.651	68.4%	85.8%	31.4%	0.879
80	0.720	76.2%	89.7%	41.5%	0.909
90	0.782	82.7%	92.9%	52.0%	0.934
100 (fully AI)	0.828	87.0%	95.0%	60.6%	0.950

Figure 1 shows the reason the response is a proportion. Half-rewritten documents do not form a third peak between the two classes. Their scores spread almost uniformly across the entire scale, so any threshold cuts the group roughly in proportion to how far the threshold sits along the score range.

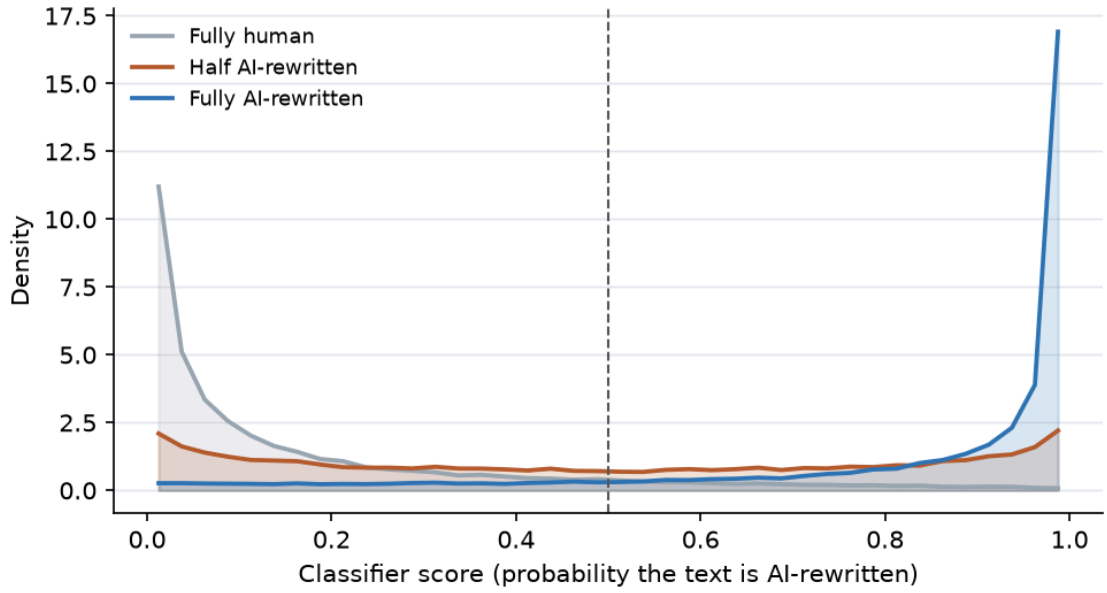

Figure 1. Classifier score distributions for fully human, half AI-rewritten, and fully AI-rewritten documents, all scored out of fold by models trained on pure texts only.

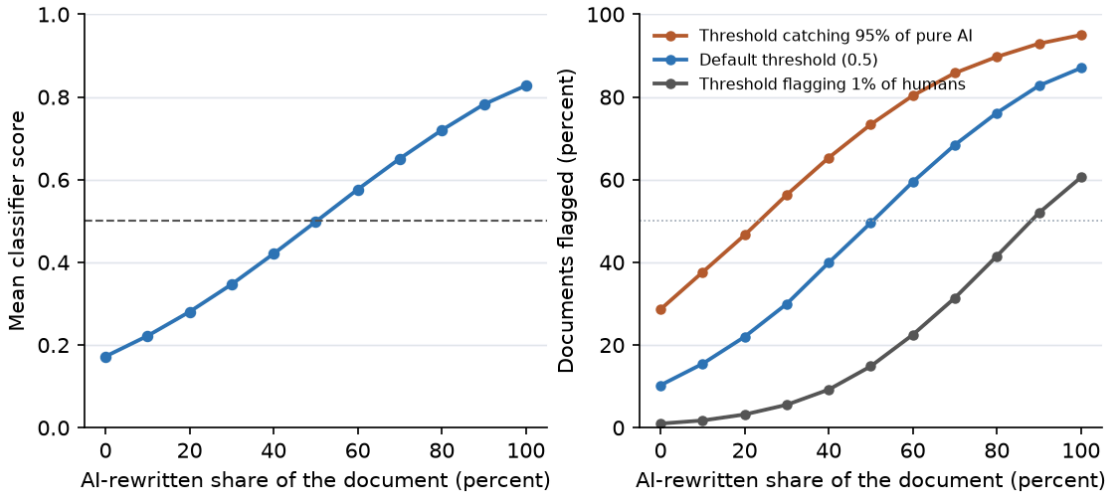

Figure 2. The classifier's response to the AI-rewritten share. Left, mean score. Right, share of documents flagged at the three fixed thresholds.

4.2 Small shares are close to invisible

The last column of Table 1 and Figure 3 measure how separable mixed documents are from fully human ones at all, regardless of threshold. A document with 10 percent of its words rewritten yields an AUC of 0.565 against fully human documents, and 20 percent yields 0.630. For comparison, random guessing yields 0.5. Below roughly a quarter of the document, AI rewriting is close to undetectable at the document level with these features, and this is not a threshold artifact but an absence of separation.

The strict threshold makes the practical case sharper. Under a one percent false positive budget, the kind of operating point an institution would need before acting on a flag [1], only 5.5 percent of documents with a 30 percent AI share are flagged, 14.9 percent at half, and barely half at a 90 percent share. Conservative document-level detection and partial rewriting are, on these numbers, almost mutually exclusive.

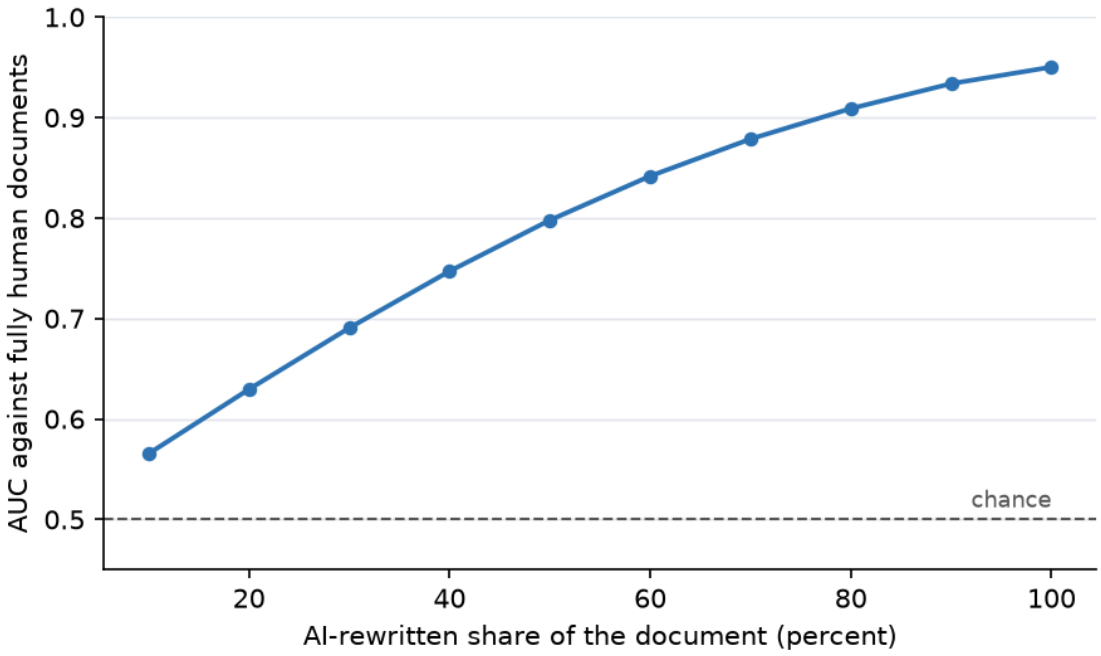

Figure 3. AUC of mixed documents against fully human documents, per AI share. The dashed line marks chance.

4.3 Position of the rewritten block

Prefix documents, in which the AI-rewritten block opens the document, score slightly but consistently higher than suffix documents at the same achieved share. At half AI content the achieved shares are 47.9 percent in both modes, yet prefix documents average 0.535 against 0.498 for suffix documents, and are flagged at 54.3 against 49.6 percent. The two halves of a rewrite therefore do not carry the AI signature equally, and documents whose opening was rewritten read as slightly more AI to the classifier. We note the effect and do not investigate the position question further, since it belongs to boundary-detection designs [4, 5].

5. How mixed documents mix

5.1 Linear in everything except sentence variation

Since almost all features are per-word or per-sentence rates, a spliced document's feature value is close to a weighted average of its parents, and the measured feature means confirm this. At half AI

share, 44 of the 45 features sit within 0.09 standard deviations of the value predicted by linear interpolation between the pure classes. The exception is sentence-length variation, which is 0.18 standard deviations above the interpolated value (Figure 4).

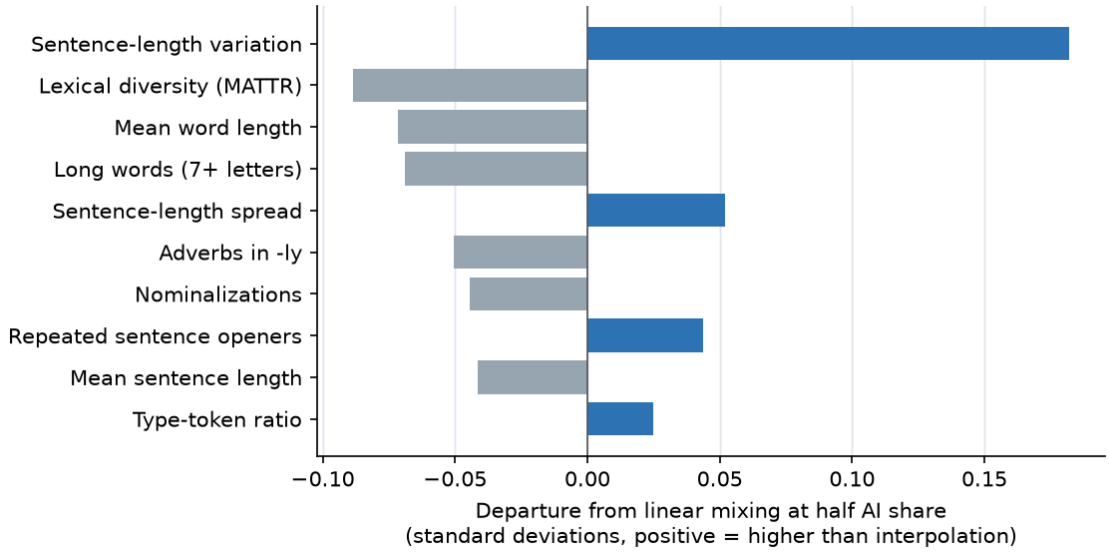

Figure 4. Departure of mixed documents from linear feature mixing at half AI share. Every feature except the sentence-length variation family mixes close to linearly.

Human academic writing has high sentence-length variation and AI rewriting compresses it [8]. A mixed document contains one block of each, and the difference in sentence rhythm between the blocks adds variation beyond what either block contains. The effect is strong enough that 81 percent of half-AI-rewritten documents show more sentence-length variation than their fully rewritten variant, and 36 percent show more than their fully human variant. At a 20 percent share, 41 percent of mixed documents exceed both variants (Figure 5).

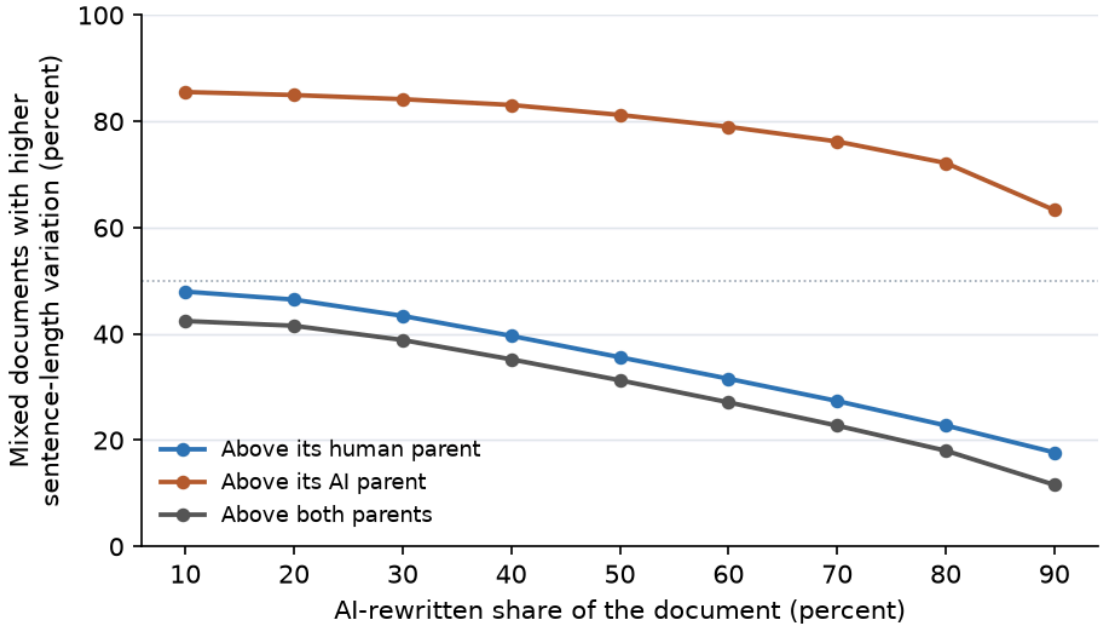

Figure 5. The share of mixed documents whose sentence-length variation exceeds that of their human parent, their AI parent, and both, per AI share.

Sentence-length variation is a human-direction marker in this classifier and in burstiness-based detection generally [8]. Splicing AI text into a human document partially restores the variation that full rewriting removes, so on this one channel, partial AI rewriting actively resembles human writing more than full AI rewriting does. Detectors that lean on burstiness are structurally weakest against exactly the mixed documents that are most common in real use.

5.2 Mixedness is barely a category

Can a classifier recognize that a document is mixed, rather than merely score it in between? We trained the same two model types to separate half-rewritten documents from pure documents of both classes, under the same grouped protocol. The linear model reaches an AUC of 0.592, close to chance, which is expected since a mixed document's features are near-averages of the pure classes and no linear boundary separates the middle of a line from its ends. The gradient boosting model reaches an AUC of 0.763 and an accuracy of 67.2 percent, finding the non-linear signature, namely, intermediate rates combined with elevated sentence variation, but at 75.0 percent recall it still mislabels 36.8 percent of pure documents as mixed. Partial rewriting is therefore detectable mainly as a lower score, and only weakly as a recognizable category of its own.

6. Differences between AI models

Figure 6 shows the response curve separately for the eight configurations. The ordering follows the pure-text detectability established in the prior studies [1, 2], and the distribution is large at every share. At half AI content, 24.9 percent of documents rewritten by Qwen are flagged, against 59.2

percent for the Gemini Flash Lite configuration. Reading the curves horizontally gives the more striking statement. A document 100 percent rewritten by Qwen is flagged at 64.5 percent, roughly the same rate as a document only 60 percent rewritten by DeepSeek. The AI share a detector can find depends on which model did the rewriting, by a factor of about two across the configurations tested.

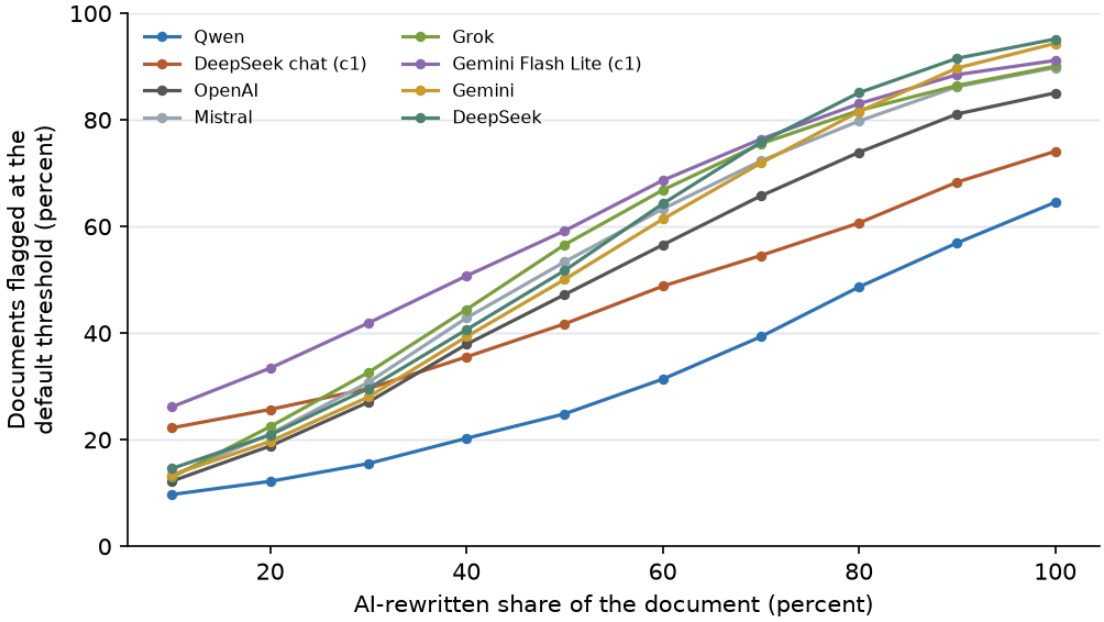

Figure 6. Share of documents flagged at the default threshold, per AI share and model configuration.

7. Interleaved rewriting

The block splices above concentrate the AI text in one place. The more common pattern in real use is scattered. Students and other writers routinely pass their drafts through AI-assisted paraphrasing tools such as ProofreaderPro, Grammarly, or QuillBot, and these tools rewrite individual sentences inside otherwise human text. The output of such a session is AI text interleaved with human text throughout the document. To measure this case, we build one further document per pair in which every second sentence of the human text is replaced by the sentence of the AI rewrite at the matching position. The AI content is then spread across the whole document instead of forming one block. The mean achieved AI share is 45.0 percent, slightly under half, since AI sentences are on average shorter than the human sentences they replace [8].

Interleaving changes nothing that the AI share does not already explain. The 25,560 interleaved documents receive a mean score of 0.478 and are flagged at 47.3 percent at the default threshold, against 0.498 and 49.6 percent for the single-block documents at half AI content. The small gap is fully accounted for by the slightly lower achieved share. The mean score of the interleaved documents lies within 0.01 of the value the block response curve predicts at a 45 percent share, and

their AUC against fully human documents is 0.780, again on the curve. Under the strict one percent false positive budget, 14.1 percent of interleaved documents are flagged. A document whose sentences have been AI-paraphrased in alternation is therefore exactly as hard to detect as a document with one rewritten block of the same total size (Figure 7, left).

Sentence-length variation remains elevated under interleaving. The interleaved documents sit 0.13 standard deviations above the value linear feature mixing predicts, and 36.1 percent show more sentence-length variation than their fully human variant, almost identical to the single block at the same share (Figure 7, right). The many short alternations do not add more variation than the single splice does. The elevation is in fact slightly smaller, 0.13 against 0.18 standard deviations, so interleaving preserves the human-direction burstiness signal of Section 5.1 without strengthening it. The practical conclusion carries over unchanged, namely, a detector reads the amount of AI text in a sentence-by-sentence paraphrased document, not the pattern, and moderate AI use through paraphrasing tools stays below the level that document-level detection can reliably find.

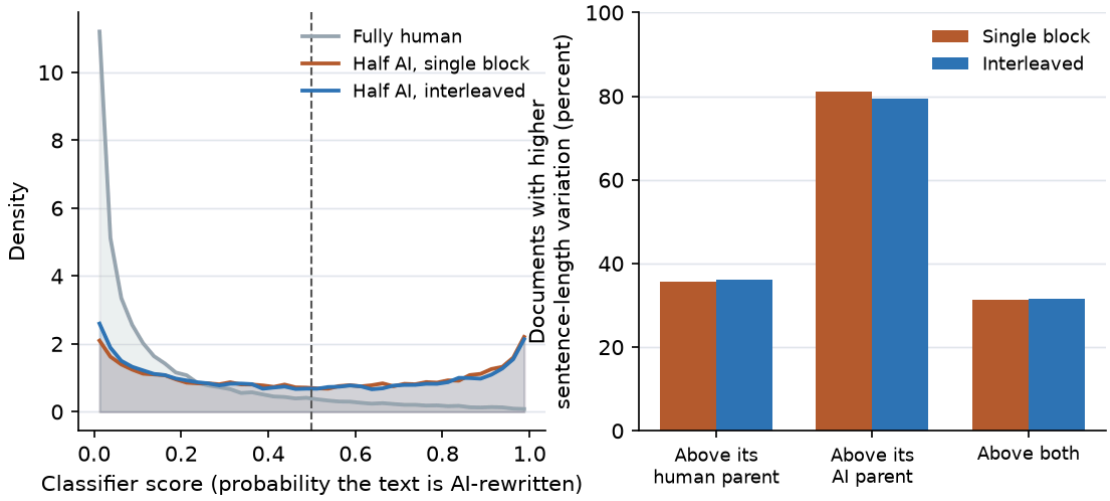

Figure 7. Interleaved against single-block mixing at approximately half AI share. Left, classifier score distributions for fully human, single-block, and interleaved documents. Right, the share of documents whose sentence-length variation exceeds that of their fully human variant, their fully AI variant, and both.

8. Discussion

The document-level score of a stylometric classifier on a mixed document is a proportion estimate, and not a full verdict. It combines two factors it cannot separate, namely, how much of the document was rewritten and how strongly the AI rewriting model marks its output. A score of 0.6 is consistent with a document fully rewritten by a model that changes its input little and with a document 60 percent rewritten by a model that marks its output strongly. Our detector-agreement study found commercial detectors giving hybrid texts the full range of verdicts [3], and the response curves here show that behavior to be a property of the mixture, not a defect of any one

tool. The 30 half-and-half hybrids of that study, built from fully generated rather than rewritten text, received a mean score of 0.57 from this classifier with 17 of 30 flagged, close to the 0.498 and 49.6 percent of the spliced half-rewrites here.

For anyone acting on detector output, the asymmetry between the two threshold regimes matters most. Aggressive thresholds flag 37.6 percent of documents that are only 10 percent AI, while also flagging 28.6 percent of fully human ones, and so cannot distinguish light AI use from no AI use at all. Conservative thresholds, the only ones defensible for individual decisions given the base rate arithmetic of our prior study [1], miss five of six half-rewritten documents. There is no threshold at which document-level stylometry reliably catches partial rewriting without heavily flagging human writing. Detecting where a document changes, rather than scoring the whole document, is the appropriate formulation for mixed texts [4, 5, 6], and our results quantify what the whole-document shortcut costs.

9. Limitations and conclusion

The spliced documents approximate partial rewriting with a single contiguous rewritten block or a strict sentence-by-sentence alternation of a content-aligned pair, and the sentence-boundary joins can locally repeat or compress content at the splice points. Documents mixed by human writers may distribute the AI content differently, though the position check and the interleaving experiment of Section 7 suggest the arrangement changes the response only modestly. All texts are English academic content of at least 300 words, the rewrites come from eight configurations captured in 2026, and the classifier is a research instrument, so nothing here describes the internals of any commercial detector, although the mixture logic applies to any detector whose score aggregates over the whole document.

A stylometric classifier reads partial AI rewriting in proportion to its amount, almost exactly linearly, with no share at which detection switches from absent to present. Shares below about a quarter of a document are effectively invisible at the document level, half-rewritten documents receive scores spread across the whole scale and are flagged at almost exactly half, and the one non-linear feature response, elevated sentence-length variation, makes mixed documents resemble human writing on the same feature burstiness-based detection relies on. Interleaved and block-spliced documents score alike at the same share, so the response depends on the amount of AI text and not on its arrangement. In brief, under a quarter of AI content the classifier cannot see the AI. Over a quarter, it sees the document as increasingly AI, almost linearly with the AI share. Whole-document AI scores on possibly-mixed documents should be read as estimates of extent of AI-ness, but never as evidence of presence.

Data availability

Per-document classifier scores for all 382,993 documents (opaque pair identifiers, corpus, model, discipline, mixing mode, target and achieved AI share, and the scores of both models), the per-

share metrics, and all analysis and figure code are openly available on Zenodo at https://doi.org/10.5281/zenodo.22054864.

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] TextPulse Research (2026). Text length and the reliability of human versus AI text classification in academic writing. Working paper. https://textpulse.ai/research/ai-detection-text-length. https://doi.org/10.5281/zenodo.22050345

[3] TextPulse Research (2026). Do AI detectors agree? An inter-rater reliability study of commercial AI text detectors on academic writing. Working paper. https://textpulse.ai/research/ai-detector-agreement. https://doi.org/10.5281/zenodo.22002869

[4] Dugan, L., Ippolito, D., Kirubarajan, A., Shi, S., and Callison-Burch, C. (2023). Real or fake text? Investigating human ability to detect boundaries between human-written and machine-generated text. Proceedings of the AAAI Conference on Artificial Intelligence, 37(11), 12763-12771. https://ojs.aaai.org/index.php/AAAI/article/view/26501

[5] Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., Afzal, O. M., Mahmoud, T., Puccetti, G., Arnold, T., Whitehouse, C., Aji, A. F., Habash, N., Gurevych, I., and Nakov, P. (2024). SemEval-2024 Task 8: Multidomain, multimodel and multilingual machine-generated text detection. Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024). https://arxiv.org/abs/2404.14183

[6] 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

[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] 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