Human versus AI Text Classification from Stylometric Features Across 121,092 Academic Texts

TextPulse Research. Working Paper.

Abstract

This study treats the separation of human-written and AI-rewritten academic text as a plain text classification task. The data are 60,306 human-written academic texts and 60,786 AI rewrites of those texts, produced by eight AI model configurations across six model families, each text described by 47 interpretable stylometric features such as word length, passive voice, lexical diversity, punctuation rates, and sentence openers. A logistic regression on these features separates the two classes with an AUC of 0.936 and an accuracy of 86.4 percent under grouped five-fold cross-validation, and a gradient boosting model increases this to an AUC of 0.961 and an accuracy of 89.5 percent. Most of the distinction between human and AI text is therefore carried by simple, countable linguistic properties of the writing. The errors have a clear structure. Identifying 99 percent of AI rewrites requires falsely flagging 68 percent of human texts, and keeping false flags at 1 percent detects only 47 percent of the rewrites. When 10 percent of a text population is AI, fewer than half of the flagged texts at the default threshold are actually AI. The misclassified human texts are the most formal ones, with longer words, more passive voice, and higher lexical diversity, and the false positive rate varies almost four times across academic disciplines. Applied without modification to an independent corpus of fully AI generated and hybrid texts, the classifier identifies all 30 generated texts, passes 26 of 30 human texts, and correlates at 0.70 with the mean verdict of nine commercial detectors. Per-text classifier scores and all analysis code are made publicly available.

1. Introduction

Whether a given text was written by a person or produced with an AI language model is now a question asked daily in classrooms, journals, and hiring processes. The commercial tools that answer it return a score or a “human written” or “AI generated” verdict, but they do not disclose what the verdict rests on, and independent evaluations report accuracies that vary widely between tools and between text types [1, 2, 3]. Given a fixed set of measurable style properties, how separable are human and AI text as classes, and what do the unavoidable errors look like?

This paper answers that question for AI rewriting of academic text. It is built on one paired corpus, in which 60,786 human-written academic texts were each rewritten by one of eight AI model configurations across six model families. Our prior studies measured what rewriting does to vocabulary [8], to sentence-length burstiness [9], and to the wider stylometric profile including

punctuation, syntax, and voice markers [10]. Each of those studies reported large paired changes. This study asks what those changes add up to. It pools all 47 usable stylometric features into a single supervised binary classification task, human versus AI, and reports the accuracy, errors, and the transfer behavior of the resulting classifier.

Since every feature in the model is an interpretable rate that a reader could count by hand, the resulting accuracy states how much of the human-AI distinction is in surface style alone, with no black-box representation involved. The gap between the interpretable model and a stronger non-linear reference model measures how much signal simple features miss.

This leads to three key results. First, the classes are highly separable at scale. The interpretable model reaches an AUC of 0.936, the non-linear reference reaches 0.961, and the small gap shows that the separation is on simple properties rather than on hidden interactions. Second, high average accuracy does not produce reliable individual verdicts. The score distributions of the two classes overlap, so every threshold trades one error for the other, and at realistic base rates most flagged texts can be human even when the classifier is right about most texts. Third, the errors are not random. The human texts that get misclassified are the most formal ones, and the AI rewrites that are missed come from the models that change the human-written source text the least.

2. Related work

Independent evaluations of commercial AI text detectors report significant error at the level of individual verdicts. Weber-Wulff et al. tested 14 detection tools on human-written, machine-translated, AI-generated, and obfuscated documents and found that no tool exceeded 80 percent accuracy, that only five exceeded 70 percent, and that content obfuscation degraded detection significantly [1]. Liang et al. showed that seven widely used detectors flagged 61 percent of essays written by non-native English speakers as AI-generated on average, while the same tools were nearly perfect on essays by native-speaking eighth-grade students, which established that detector false positives concentrate on specific groups of human writers [2]. Our prior study measured agreement between nine commercial detectors on 90 academic texts and found that the verdict a text receives depends heavily on which tool is used [7].

On the theoretical side, Sadasivan et al. constrained the performance of any possible detector by the total variation distance between the human and AI text distributions, and showed empirically that paraphrasing degrades a wide range of detection schemes [3]. Their result formalizes the intuition that detection accuracy is a property of the overlap between two distributions, not of any particular tool. The present study measures that overlap directly for one controlled setting, using fixed interpretable features.

Feature-based classification of machine-generated text predates the current generation of AI models. Fröhling and Zubiaga built classifiers on hand-crafted linguistic features to detect GPT-2, GPT-3, and Grover output, and reported performance competitive with far more expensive methods [4]. Ippolito et al. showed that automatic classifiers and human raters make different

errors on generated text, and that classifier accuracy depends heavily on the decoding strategy of the AI generator [5]. Muñoz-Ortiz et al. documented systematic linguistic differences between human news text and the output of six AI models [6]. The present study differs from these in its paired design. Every AI text is a rewrite of a specific human text in the same corpus, so the two classes are matched on topic, discipline, and content, and the classifier cannot succeed by learning the subject matter of the texts.

The feature set itself comes from our prior work. The vocabulary study established per-family word injection and deletion patterns [8], the burstiness study established that all eight configurations compress sentence-length variation [9], and the stylometric study measured the full feature set used here and demonstrated that the same features identify which model produced a rewrite at 3.9 times chance accuracy [10].

3. Data and methods

3.1 Corpus

The corpus is the same collection of human-AI paraphrase pairs used in the three prior studies, built from open-access human written academic text [8, 9, 10]. Corpus 1 comprises 38,323 pairs rewritten by two production configurations, namely a DeepSeek chat configuration and a Gemini Flash Lite configuration. Corpus 2 comprises 22,463 pairs rewritten by six models, namely DeepSeek, Gemini, Grok, Mistral, OpenAI, and Qwen. Together, they provide 60,786 pairs in total. Human sources average 343 words and AI rewrites 321 words.

A small number of human source texts appear in more than one pair, namely 402 sources with two pairs and 39 with three. To prevent any source from influencing both training and evaluation, we grouped all texts by their human source, deduplicated the human side to one instance per source, and performed every split at the group level. The classification table therefore holds 121,092 texts, of which 60,306 are distinct human texts and 60,786 are AI rewrites.

3.2 Features

Each text is described by the stylometric feature set defined in the prior study [10]. The features cover word and sentence counts, mean word length, the rate of words with seven or more letters, type-token ratio, moving-average type-token ratio, sentence-length mean, standard deviation, and coefficient of variation, punctuation rates per 1,000 words, passive voice, nominalizations, first person, sentence-opener repetition, connective sentence openers, contractions, question and exclamation marks, adverbs ending in -ly, and the rates of 25 hedge and connective phrases. We use 47 of the 49 features. The em dash and en dash rates are excluded, since the prior study restricted those two features to comparisons between AI models only, and not between human and AI texts [10]. Passive voice, nominalization, and the -ly adverb rate are pattern-based approximations, computed identically for both classes.

3.3 Classification protocol

The primary model is a logistic regression on standardized features (scikit-learn 1.9.0, L2 regularization, C = 1.0), chosen because its coefficients can be read directly. A histogram gradient boosting classifier trained on the same splits serves as a stronger non-linear reference. Both models use balanced class weights.

All results come from five-fold cross-validation grouped by human source, so no source text or any rewrite of it appears in both training and test data. Scores are pooled out-of-fold, which gives every one of the 121,092 texts a score from a model that never saw its source. Reported accuracies use the default 0.5 threshold unless stated otherwise. As a robustness check we removed the two raw length features, words and sentence count, and reran the full protocol. The AUC of the logistic regression changed from 0.936 to 0.934 and that of the gradient boosting model from 0.961 to 0.958, so the classification does not rest on text length.

3.4 Transfer corpus

To test behavior outside the training distribution, we applied the trained logistic regression, without any modification, to the 90 texts of our detector-agreement study [7]. That corpus contains 30 verifiably pre-2022 human-written academic texts, 30 fully AI-generated texts from five model families, and 30 hybrid texts splicing human and AI halves. The texts were scored by nine commercial detectors in the prior study, which allows a direct comparison between this paper's 47-feature classifier and commercial systems on the same texts. The classifier was trained only on rewrites of academic text, while the transfer corpus contains fully generated text produced from instructions.

4. Classification results

4.1 Overall accuracy

The logistic regression separates human texts from AI rewrites with a pooled AUC of 0.936 and an accuracy of 86.4 percent. Per-fold AUC ranges from 0.934 to 0.939, so the result is stable across splits. The gradient boosting reference reaches an AUC of 0.961 and an accuracy of 89.5 percent. Table 1 summarizes both models at the default threshold.

Model	AUC	Accuracy	AI rewrites identified	Human texts falsely flagged	Precision
Logistic regression	0.936	86.4%	86.0%	13.2%	86.8%
Gradient boosting	0.961	89.5%	88.8%	9.8%	90.1%

The gap between the interpretable model and the non-linear reference is 2.5 points of AUC and about 3 points of accuracy. Most of the separation between human and AI academic text is therefore carried by simple and individually countable rates, and only a small remainder requires feature interactions that a linear model cannot represent. The same pattern appeared in the attribution experiment of the prior study [10].

Figure 1 shows the pooled score distributions. Both classes concentrate near their correct ends of the scale, and both leave a long tail in the middle. This overlap region produces every error, and no threshold placement removes it.

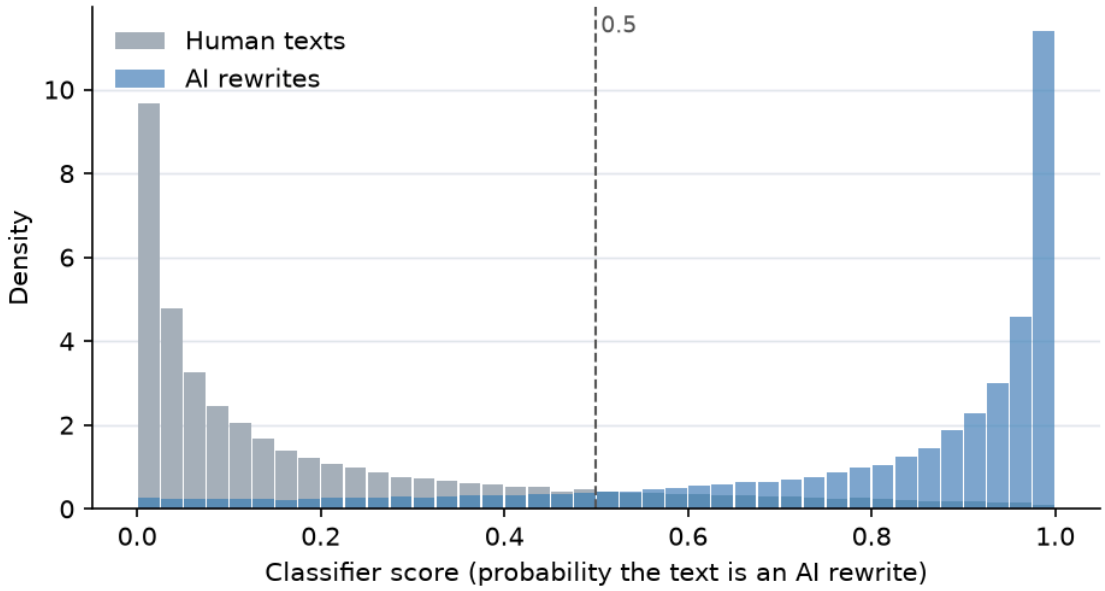

Figure 1. Classifier score distributions for human texts and AI rewrites, pooled out-of-fold scores from grouped five-fold cross-validation.

4.2 What carries the signal

The largest standardized coefficients are in the directions established by the paired measurements of the prior studies. Passive voice, the rate of long words, mean word length, lexical diversity, and the -ly adverb rate push a text toward the AI class. High sentence-length variation, repeated sentence openers, and connective sentence openers push a text toward the human class. Among the phrase features, “furthermore”, “underscore”, and “pivotal” carry the most weight, consistent with the injection rates measured before [8, 10]. The classifier works because AI rewriting changes many simple properties at once, in one direction, and across every model tested.

4.3 Transfer between corpora

Training on corpus 1 alone and testing on corpus 2 gives an AUC of 0.861 and an accuracy of 74.0 percent. The reverse direction gives an AUC of 0.808 and an accuracy of 73.1 percent. Accuracy against unseen model configurations therefore drops by roughly 12 to 13 points against the pooled

result. A classifier of this kind meets new models constantly in practice, so the pooled accuracy of 86 to 90 percent should be read as specific to the configurations it was trained on, and the transfer numbers as the more realistic expectation for unseen ones.

5. The structure of the errors

5.1 Every threshold trades one error for the other

Since the two score distributions overlap, the threshold sets an exchange rate between the two error types rather than an error level. Table 2 gives the operating points of the logistic regression.

Target	AI rewrites identified	Human texts falsely flagged
Identify 90% of AI	90.0%	18.8%
Identify 95% of AI	95.0%	32.1%
Identify 99% of AI	99.0%	68.0%
Flag at most 1% of humans	46.9%	1.0%
Flag at most 5% of humans	72.4%	5.0%
Flag at most 10% of humans	82.5%	10.0%

The two ends of the table state the problem explicitly. A policy that insists on finding 99 percent of AI rewrites falsely flags 68 percent of human texts. A policy that protects human writers by flagging at most 1 percent of them finds less than half of the AI rewrites. Figure 2 traces the full curve between these points.

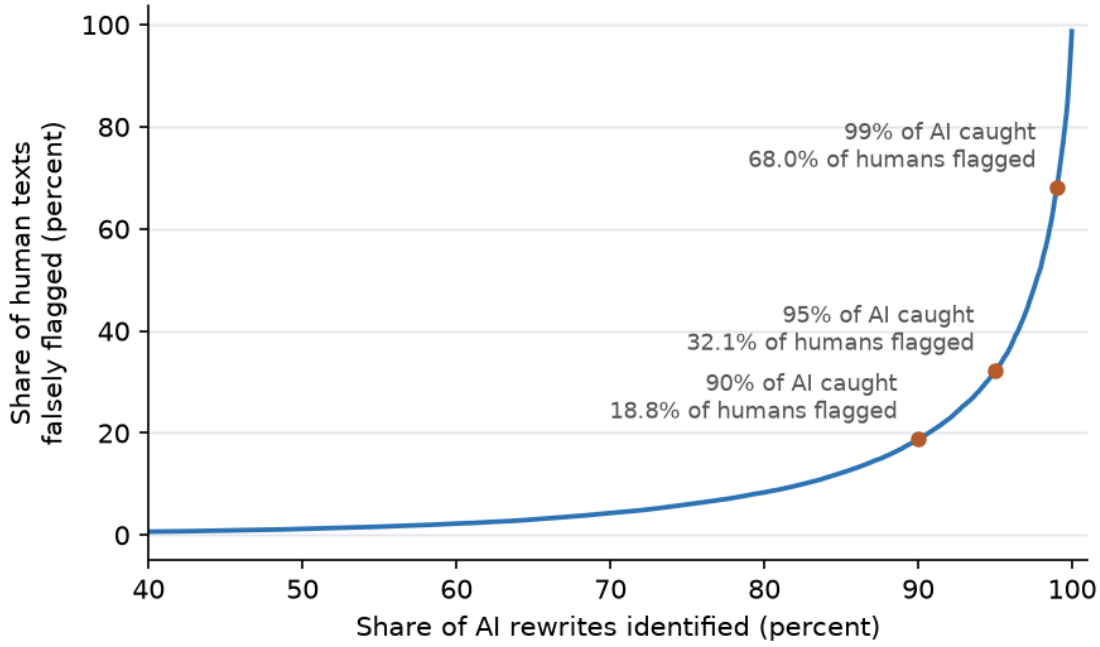

Figure 2. The share of human texts falsely flagged at each level of AI identification. Marked points show the cost of identifying 90, 95, and 99 percent of AI rewrites.

5.2 Base rates decide what a flag means

Accuracy is a property of the classifier, but the meaning of an individual flag depends on how much AI text the population actually contains. When AI rewrites are 10 percent of a population, the default threshold produces flags of which only 41.9 percent are truly AI, since the 13.2 percent false positive rate applies to the much larger human majority. At a 5 percent base rate the share of correct flags drops to 25.5 percent, and at 1 percent it drops to 6.2 percent, namely roughly fifteen wrong flags for every right flag. A strict threshold that flags only 1 percent of humans increases the share of correct flags to 83.9 percent at a 10 percent base rate, but it identifies less than half of the AI texts. Figure 3 shows the full curves. These numbers are arithmetic consequences of the error rates in Table 2, and they apply with equal impact to any detector with comparable error rates, including commercial ones [1, 7].

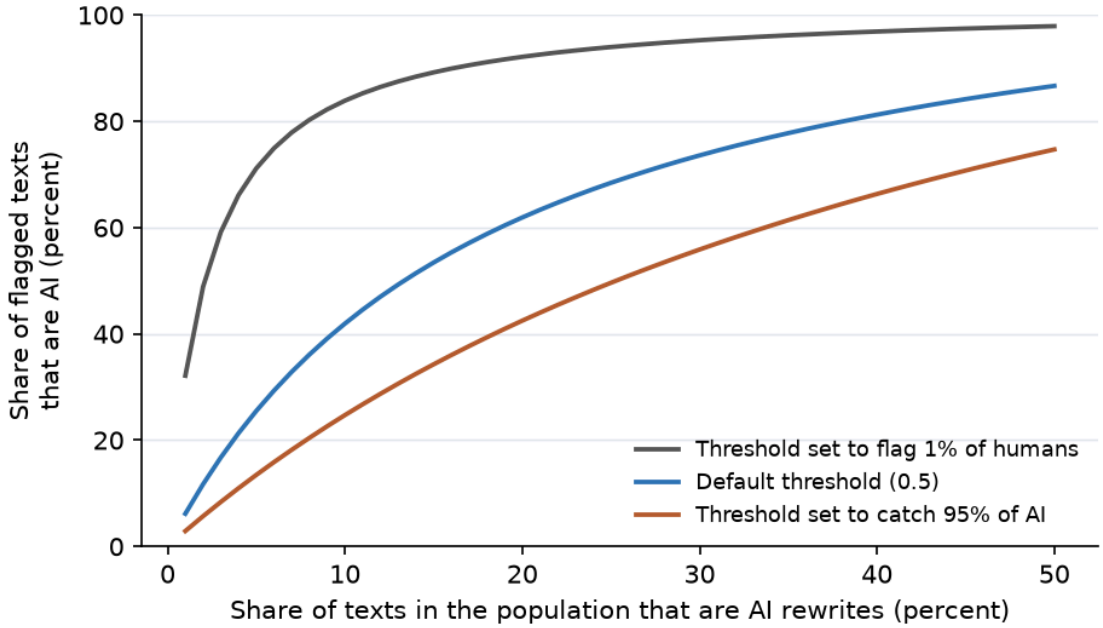

Figure 3. The share of flagged texts that are actually AI, as a function of the AI share of the population, for three threshold policies.

5.3 Which human texts get misclassified

The 7,977 human texts flagged at the default threshold are not a random sample. Compared with correctly classified human texts, they average 5.89 versus 5.45 characters of word length (a gap of 1.14 within-class standard deviations), 412 versus 348 long words per 1,000 words, 20.9 versus 13.4 passive constructions per 1,000 words, higher type-token ratio and higher moving-average lexical diversity, more -ly adverbs, more nominalizations, and more commas. They also repeat sentence openers less than half as often and show lower sentence-length variation. Figure 4 shows the profile.

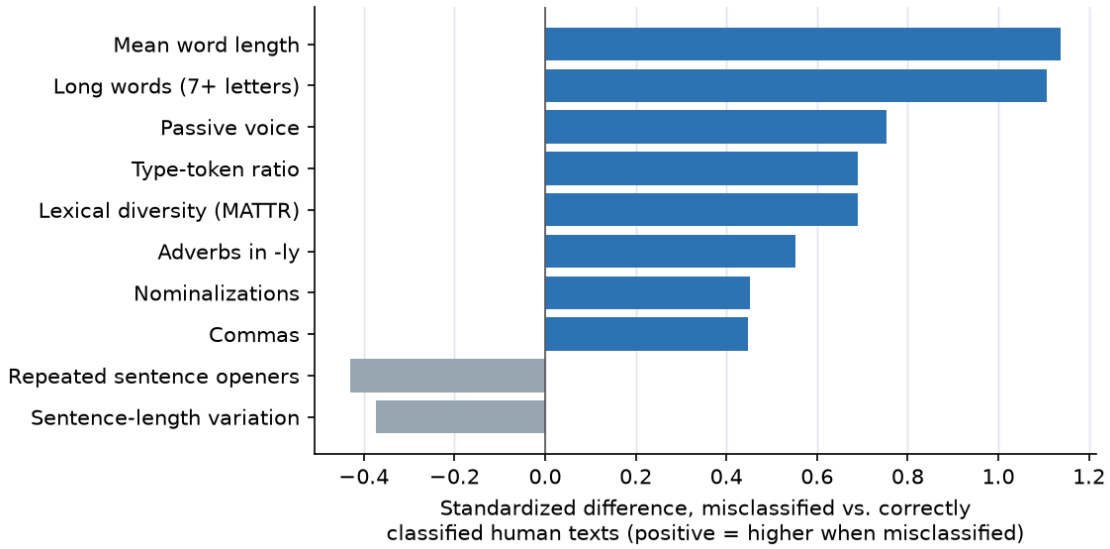

Figure 4. Feature profile of misclassified human texts, expressed as standardized differences against correctly classified human texts.

This is the profile of well-written formal content. The human texts that read most like AI rewrites are the ones whose authors use precise long vocabulary, write in the passive voice of their field, vary their word choice, and keep their sentence rhythm flat. Our prior studies demonstrated that AI rewriting moves every text in this direction [9, 10], so the writers whose natural register already matches that direction receive the false positives the most.

The effect is visible across disciplines. In corpus 1, the false positive rate at the default threshold ranges from 8.2 percent in economics and finance and 8.7 percent in law and humanities to 22.9 percent in biology and chemistry, 27.8 percent in physics and materials science, and 30.4 percent in engineering. In corpus 2 the rates are lower overall but preserve a similar ordering, from 5.8 percent in sociology to 11.7 percent in psychology. A discipline's default writing style exposes it to false flags. This is in line with the finding of Liang et al. that detector false positives concentrate on specific writer groups [2], with the concentration here falling on the more technical fields (e.g., engineering).

The same pattern appears in overall accuracy. Figure 5 shows the combined classification accuracy by discipline at the default threshold, which ranges from 80.4 percent in biology and chemistry to 89.9 percent in sociology. The technical and laboratory fields occupy the lower end in both corpora, and the social sciences, economics, and humanities occupy the upper end.

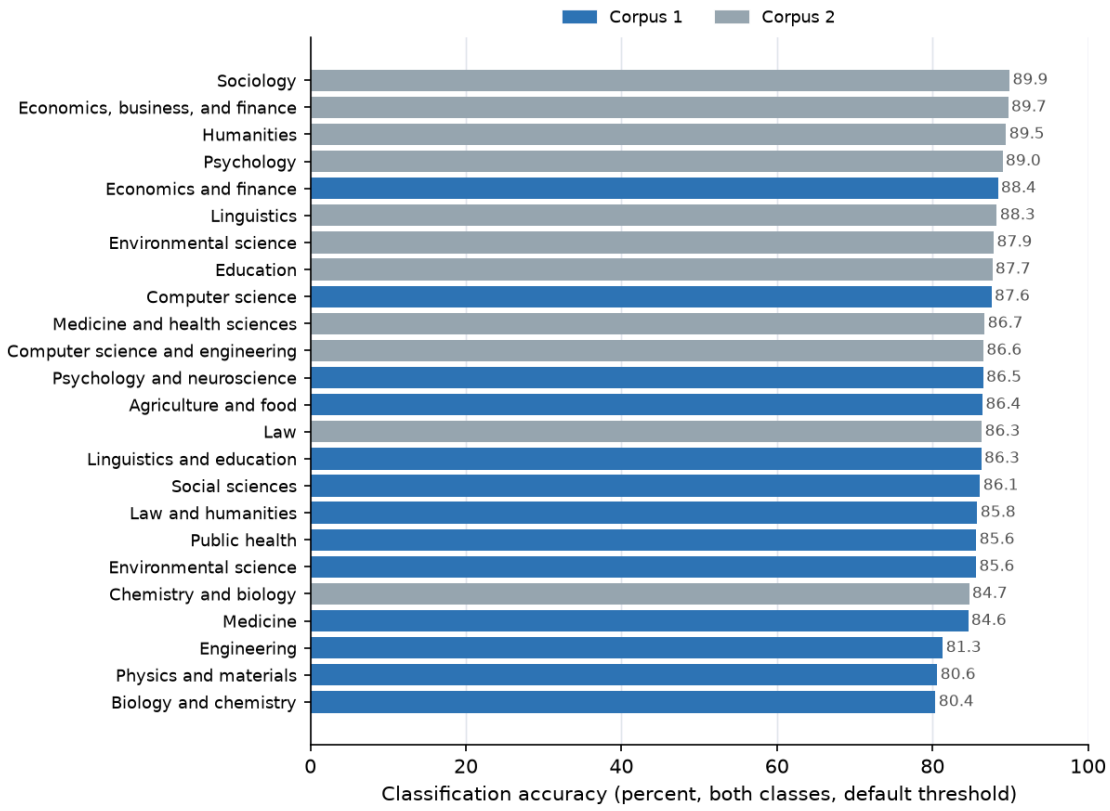

Figure 5. Classification accuracy by discipline at the default threshold, both classes combined, colored by corpus.

5.4 Which AI rewrites are missed

The share of AI rewrites classified as human ranges from 7.3 percent for DeepSeek and 7.6 percent for Gemini to 20.7 percent for OpenAI and 44.5 percent for Qwen. The ordering matches the style-distance ranking of the prior study, in which Qwen changed the style of the source text least and DeepSeek changed it most [10], with a single adjacent transposition of Grok and Mistral. Figure 6 presents all eight configurations.

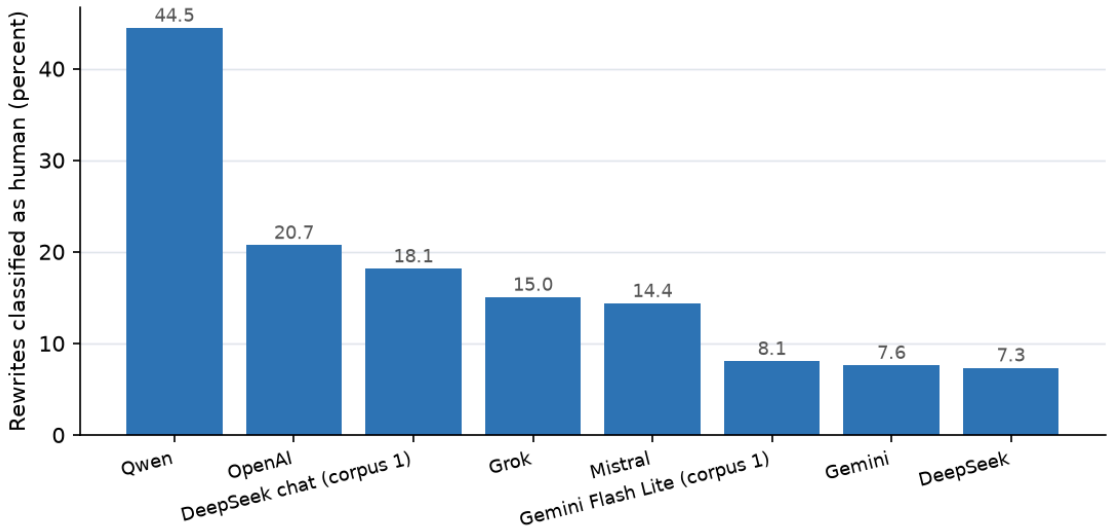

Figure 6. The share of each configuration's rewrites classified as human at the default threshold.

The practical reading is direct. A model that intervenes lightly in a human text produces a rewrite that stays inside the human score distribution, and almost half of Qwen's rewrites do. Detectability is not a property of AI text in general. It is a property of how much a particular configuration changes the input, and the configurations differ by a factor of six.

6. Comparison with commercial detectors

Applied unchanged to the 90 texts of the detector-agreement corpus, the classifier produces verdicts closely in line with the nine commercial detectors scored there, although it was never trained on this corpus or this task. It assigns the 30 human texts a mean score of 0.26 and keeps 26 of them below the 0.5 threshold. It assigns the 30 AI generated texts a mean score of 0.91 and places all 30 above the threshold. On the 60 unambiguous texts its verdicts agree with the individual commercial detectors between 85.0 and 93.3 percent of the time, and its ranking of the texts correlates with the mean score of the nine detectors at a Spearman correlation of 0.70, with per-detector correlations from 0.57 (Winston) to 0.75 (GPTZero). The 30 hybrid texts are in between, with a mean score of 0.57 and 17 of 30 flagged, which mirrors the intermediate and inconsistent treatment hybrids received from the commercial AI detector tools [7]. Figure 7 summarizes the comparison.

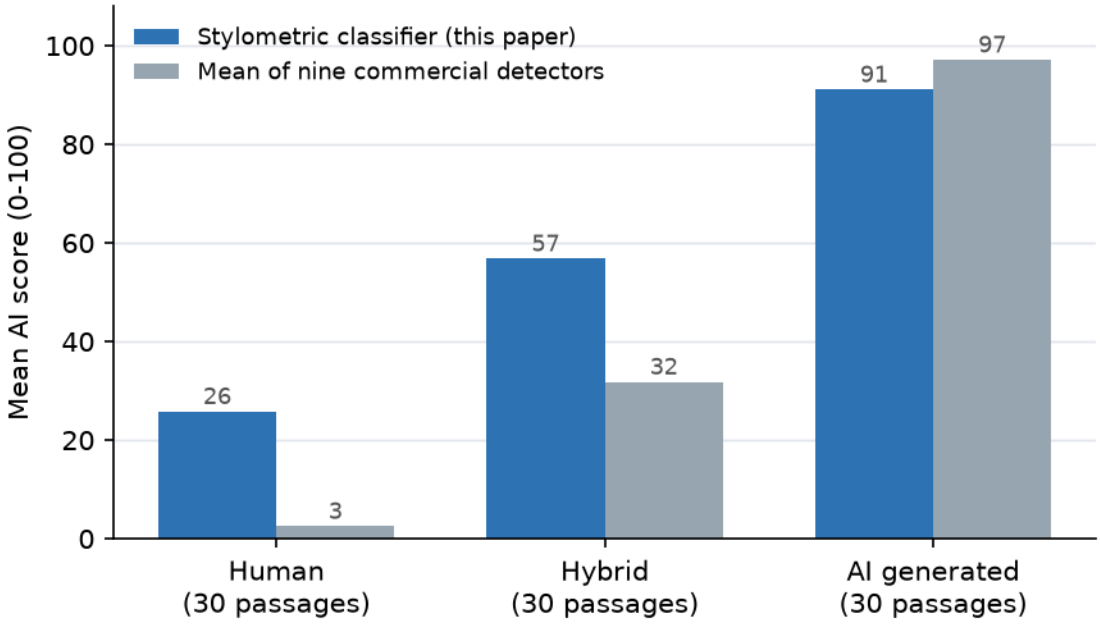

Figure 7. Mean AI score by condition on the 90-text transfer corpus, for this paper's classifier and for the mean of the nine commercial detectors scored in the prior study.

First, 47 countable surface rates reproduce a large part of the ranking behavior of commercial detection systems on an independent corpus. This suggests that most of what those systems respond to is ordinary measurable style. Second, the transfer did not remove the overlap problem. Four of 30 human texts were still flagged, and those four were the formal, technical ones, exactly as Section 5.3 predicted.

7. Discussion

Treated purely as a simplified text classification task, classifying human academic writing from AI rewriting is largely solved by surface stylometry at the population level. An interpretable model reaches an AUC of 0.936 on held-out sources, a non-linear model adds only 2.5 points, and the features that carry the signal are the same simple rates that our prior studies measured one by one. For research questions at corpus scale, namely, estimating how much AI-processed text a collection contains or tracking stylistic drift over time, classifiers of this kind are accurate, cheap, and fully auditable.

The same numbers argue against reading any single flag as a verdict about a person. The score distributions overlap, so thresholds only move errors between the two classes. The false positives concentrate on formal, technical, carefully edited human writing, and the false negatives concentrate on lightly AI rewritten text. At reasonable base rates, a majority of flagged texts can be human. None of this is a defect of one tool. Our classifier, built openly on 47 named features, shows the same error structure that independent evaluations report for commercial AI detectors such as

Turnitin and GPTZero [1, 2, 7], and theory predicts the pattern for any detector whose classes overlap [3].

The per-model result adds a time dimension. Detectability tracks how much a configuration changes its input, the configurations tested here differ six-fold on that quantity, and configurations change with every model update. A classifier tuned to today's models met unseen configurations in our own transfer experiment and lost 12 points of accuracy. Any fixed accuracy claim about AI text detection is therefore a claim about specific models at a specific point in time.

8. Limitations and conclusion

The training corpus is AI rewriting of English academic text, and the population-level results are specific to that setting. The transfer experiment shows encouraging behavior on fully AI generated text, but it covers 90 texts from one prior study. Passive voice, nominalization, and -ly adverb rates are pattern-based approximations rather than parses. The eight configurations were captured at fixed time points in 2026. The classifier is a research instrument, and nothing here implies that commercial detectors use these features or these thresholds.

The stylometric changes documented for vocabulary, burstiness, punctuation, and syntax add up to a highly separable classification problem, with an AUC near 0.96 from countable features alone. The residual overlap is not noise, but structure. It is occupied by human writers whose formal writing style resembles the direction in which AI writes text, and occupied by AI models that rewrite human text the least. Classification at this accuracy supports population-level measurement well and individual verdicts poorly, and the gap between those two uses is set by base rates and overlap, and not by the general quality of any specific detector tool.

Data availability

Per-text classifier scores for all 121,092 texts (opaque pair identifiers, corpus, model, discipline, side, and the out-of-fold scores of both models), the transfer-corpus comparison table, and all analysis and figure code are openly available on Zenodo at https://doi.org/10.5281/zenodo.22048476.

References

[1] 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

[2] 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

[3] 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

[4] Fröhling, L., and Zubiaga, A. (2021). Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. PeerJ Computer Science, 7, e443. https://doi.org/10.7717/peerj-cs.443

[5] 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/

[6] Muñoz-Ortiz, A., Gómez-Rodríguez, C., and Vilares, D. (2024). Contrasting linguistic patterns in human and LLM-generated news text. Artificial Intelligence Review, 57(9). https://arxiv.org/abs/2308.09067

[7] 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

[8] 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

[9] 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

[10] 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