Non-Native English Writing and the False Positives of Stylometric AI Text Detection

TextPulse Research. Working Paper.

Abstract

A widely cited study found that perplexity-based AI text detectors flag the majority of essays by non-native English writers as AI-generated while sparing native writers, and the finding has impacted the debate on AI detection in education. This study measures whether the bias is true for a transparent stylometric classifier. We score all 5,600 essays of the ICNALE corpus, written under controlled conditions by college students from ten Asian countries and regions at known proficiency levels and by English native speakers, with the human-versus-AI classifier of our prior studies, trained on 50,701 academic texts and rewrites and never on any essay. Every essay predates modern AI writing tools, so every flag is a false positive. The logistic regression classifier shows no bias against non-native writers. It flags 2.7 percent of learner essays and 3.3 percent of native essays at the default threshold, statistically indistinguishable rates, and both far below the 10.3 percent it flags on full-length human academic texts and the 24.3 percent on length-matched academic excerpts. The proficiency gradient runs opposite to the published bias, as the most proficient learners are flagged the most, namely, 5.0 percent at B2 against 1.9 percent at A2, since proficient writing drifts toward the formal register the classifier associates with an AI writing style. Learner limitations, namely, simple everyday vocabulary, repetitive word choice, and uneven sentence lengths, are the opposite of the AI signature under a stylometric analysis, while a perplexity (surprisal) analysis reads the same limitations as machine-like predictability. The protection is not intrinsic to the feature set, however. A gradient boosting model trained on identical features and data flags learners at 12.5 percent against 3.8 percent for natives, a threefold gap with the least proficient flagged most. Whether AI detection discriminates against non-native writers is a property of the detector itself, not of the human writing style. Per-essay scores and all analysis code are made publicly available.

1. Introduction

The strongest fairness result in the AI detection literature involves non-native English writers. Liang et al. showed that seven widely used GPT detectors flagged 61 percent of TOEFL essays by non-native speakers as AI-generated on average, while the same detectors were nearly perfect on essays by native-speaking students [1]. The mechanism they proposed is perplexity. Detectors that score a text by how predictable it is under a language model read the limited vocabulary and conventional phrasing of second-language writing as machine-like, so the writers with the least English are flagged the most. The result is cited as evidence that AI detection is structurally unfair

to the largest group of English writers in the world, namely, those who learned it as a second language.

Our prior studies built a different kind of “stylometric” detector. A logistic regression over interpretable stylometric features, namely, sentence-length statistics, vocabulary diversity, punctuation rates, and phrase markers, categorizes fully human from fully AI-rewritten academic texts with an AUC of 0.936 across 121,092 texts [2]. Its false positives are not random. They concentrate on formal, technical, edited human writing, the writing that most resembles the AI register [2], and on short texts [3]. Whether they also concentrate on non-native writers is an open question, and the direction is not obvious. A perplexity detector punishes predictable word choice. A stylometric detector punishes polished uniformity. Second-language writing is predictable in its vocabulary but rarely polished in its form. Hence, our stylometric-based detector correctly avoids flagging non-native English writing as AI, which is not the case for the perplexity-based detectors used in earlier studies.

The ICNALE Written Essays corpus contains 5,600 essays produced under controlled conditions on two fixed prompts by college students from ten Asian countries and regions, each placed in a proficiency band aligned with the CEFR scale, alongside essays by English native speakers on the same prompts [4, 5]. The corpus was collected and released years before modern AI writing tools existed, so every essay is certainly human-written and every flag is a false positive. Since natives and learners wrote the same prompts under the same conditions, the learner-native comparison is free of the topic and genre confounds that complicate most fairness evaluations.

We score every essay with the classifier of our prior studies, trained only on academic texts and their AI rewrites and never on any essay, and measure false positive rates by first language, by proficiency band, and under professional human editing. We also score the corpus with a gradient boosting model trained on identical features and data, and the two models answer the fairness question differently. The sections below quantify both answers.

2. Related work

Liang et al. established the non-native bias for perplexity-based detectors, with false positive rates above 90 percent for some tools on TOEFL essays, and showed that prompting an AI model to enrich the vocabulary of a flagged essay removed the flag [1]. Perkins et al. tested six commercial detectors on 805 texts and measured an average accuracy of 39.5 percent, falling to 17.4 percent once the machine-generated text was modified by simple evasion techniques, and framed those error rates as a question of inclusive assessment, identifying non-native English speakers as the group most exposed to false accusation [6]. Weber-Wulff et al. evaluated fourteen detection tools and found no tool exceeding 80 percent accuracy, with false positives unevenly distributed across writing styles [7]. The common thread is that an aggregate accuracy number conceals the question this study asks, namely, which human writers absorb the false positives.

Detection methods differ in what they measure, and the difference determines exactly which writers they misread. One family scores a text by how predictable it is under a reference language model. GLTR ranks each token by its probability, its rank, and the entropy of the predicted distribution, and presents the result to a human reader [8]. DetectGPT tests whether a passage occupies a region of negative curvature in a model's log probability function, a property of model-sampled text that human text lacks [9]. Both treat predictability as evidence of machine authorship, which is the mechanism Liang et al. identify as the source of the non-native penalty [1]. A second family measures surface style instead, namely, sentence-length statistics, vocabulary diversity, punctuation rates, and phrase frequencies, and checks whether a text carries the register of AI writing rather than whether a model would have predicted it [2, 10]. The two families answer different questions, so there is no reason for their false positives to fall on the same writers. This study measures where they fall for the second family.

A related line of work shows how easily detector output moves under rewriting. Sadasivan et al. showed that a light paraphraser applied on top of a generator breaks watermarking schemes, neural detectors, and zero-shot classifiers alike, and argued that the best possible detector degrades toward chance as generators approach human text [11]. Dugan et al. assembled a benchmark of more than six million generations across eleven models, eight domains, and eleven adversarial attacks, and found that detectors reporting accuracies above 99 percent are fooled by it [12]. Our splicing study found the same sensitivity from the other direction, since a document's score tracks the proportion of it that has been AI-rewritten rather than only the presence of AI rewriting [10]. The editing result reported in Section 8 involves this line and indicates that the AI rewriting need not be machine rewriting for the score to change.

Lu built an automatic system that computes fourteen indices of syntactic complexity for second-language writing and validated it on college-level learner data [13]. Crossley and McNamara predicted second-language writing proficiency from linguistic features and found that the more proficient writers produce texts that are more lexically sophisticated rather than more cohesive [14]. Both results describe the same direction of travel, namely, that as proficiency rises, learner writing acquires the longer and more even sentences and the wider vocabulary that a stylometric classifier reads as the AI register. Second-language research therefore predicts the proficiency gradient measured in Section 5, including its direction.

The ICNALE corpus family was built by Ishikawa for contrastive interlanguage analysis, with controlled prompts, conditions, and proficiency banding designed to make learner groups comparable [4, 5]. The Edited Essays module adds professional native-speaker editing of a subset of learner essays [15], which we use to test how human editing moves a detector score. To our knowledge, no prior work has scored a large learner corpus with a detector whose features, training data, and thresholds are fully known, which is what allows the mechanism behind the "fairness" result to be read off directly.

3. Data and methods

3.1 The ICNALE essays

The ICNALE Written Essays module (version 2.6) contains 5,600 essays by college students in China, Hong Kong, Indonesia, Japan, Korea, Pakistan, the Philippines, Singapore, Thailand, and Taiwan, and by English native speakers [4, 5]. Every participant wrote two argumentative essays on fixed prompts, one on part-time jobs for students and one on smoking in restaurants, under controlled conditions with a length guideline of 200 to 300 words. Learners are placed into four proficiency bands aligned with the CEFR scale, namely, A2, lower B1, upper B1, and B2 or above, based on standardized test scores. The corpus comprises 5,200 learner essays from 2,600 students and 400 native-speaker essays from 200 writers. The median essay contains 218 words by our tokenization.

The Edited Essays module (version 3.1) contains 656 of the learner essays in two forms, the original and a version corrected and made natural by professional native-speaker editors [15]. We score both forms and compare them pairwise. As a length-robustness check, we also score each participant's two essays concatenated into one document with a median of 441 words, which clears the 300-word cohort rule of our prior studies [3, 10].

3.2 Classifier and protocol

The classifier and training protocol are those of our mixed-authorship (part AI part human text) study [10]. Each text is described by 45 stylometric features, and a logistic regression on standardized features is trained on the pure texts of the series corpora, namely, 25,140 deduplicated human academic texts and 25,561 full AI rewrites of them by eight AI model configurations, under five-fold cross-validation grouped by human source. A histogram gradient boosting model is trained under the identical protocol as the non-linear reference. As validation, the pooled out-of-fold AUC on the pure texts reproduces the prior numbers exactly, 0.950 for the logistic regression and 0.968 for the gradient boosting model [3, 10].

Every ICNALE essay is out of sample for every fold, and each essay is scored by the mean probability of the five fold models. Three thresholds are evaluated, following the prior studies, namely, the default threshold of 0.5, the threshold that identifies 95 percent of AI rewrites (0.20 for the logistic regression), and the strict threshold that flags 1 percent of academic humans (0.91). For the gradient boosting model the corresponding thresholds are derived from its own scores. Confidence intervals for flag rates are 95 percent bootstrap intervals clustered by student, so the two essays of one writer are never treated as independent.

3.3 Baselines

The first baseline is the 25,140 human academic texts themselves, scored out of fold, of which 10.3 percent are flagged at the default threshold [2]. Second, the same academic texts truncated at a sentence boundary to about 230 words, the ICNALE essay length, and scored by the fold model that never saw their source. Truncation raises the academic false positive rate to 24.3 percent,

which is consistent with the short-text penalty of our length study [3]. This second baseline is the "fair" comparison for the essays, since it has both the human origin and the length constant and varies only the kind of writing.

4. Essays are flagged far less than academic prose

The logistic regression flags 2.7 percent of learner essays (95 percent CI 2.2 to 3.1) and 3.3 percent of native essays (CI 1.5 to 5.0) at the default threshold. The two rates are statistically indistinguishable, and the learner point estimate is the lower one. There is no trace of the published bias. For comparison, the same classifier flags 10.3 percent of full-length human academic texts and 24.3 percent of academic texts cut to the same length as the essays. At matched length, an essay is seven to nine times safer than academic text.

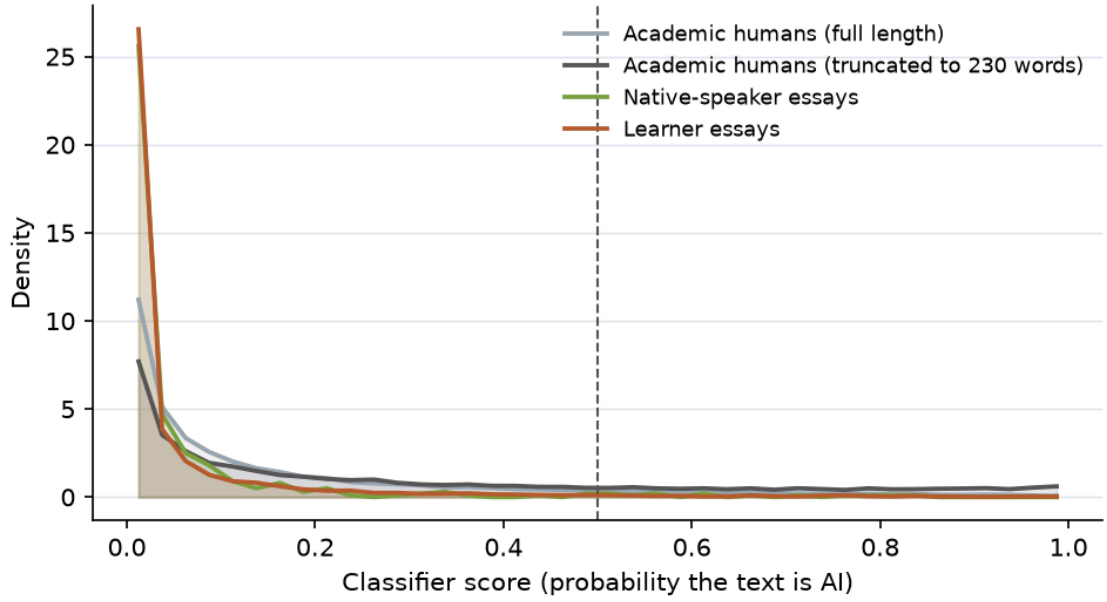

Figure 1. Classifier score distributions. Human academic texts at full length and truncated to essay length, native-speaker essays, and learner essays. Essays of both groups concentrate near zero.

The strict threshold designed to flag 1 percent of academic humans flags 0.2 percent of learner essays and no native essay at all. The permissive threshold that catches 95 percent of AI rewrites, which misflags 28.6 percent of academic humans, still flags only 8.7 percent of learner and 7.0 percent of native essays.

The ten learner groups span a narrow range, from 0.3 percent (Pakistan) and 0.6 percent (Thailand) to 3.8 percent (China, Japan) and 4.5 percent (Korea). The native speakers, at 3.3 percent, rank seventh of the eleven groups, in the middle of the learner range. The factors that separate the groups do not separate natives from learners.

The two prompts differ more than the eleven groups. The smoking essays, argued in a public-policy register, are flagged at 4.7 percent across all groups, while the part-time-job essays, argued from personal experience, are flagged at 0.7 percent. Register moves the false positive rate by a factor of more than six within the same writers, another instance of the pattern of our prior study, namely, that the classifier's false positives track formality rather than any property of the writer [2].

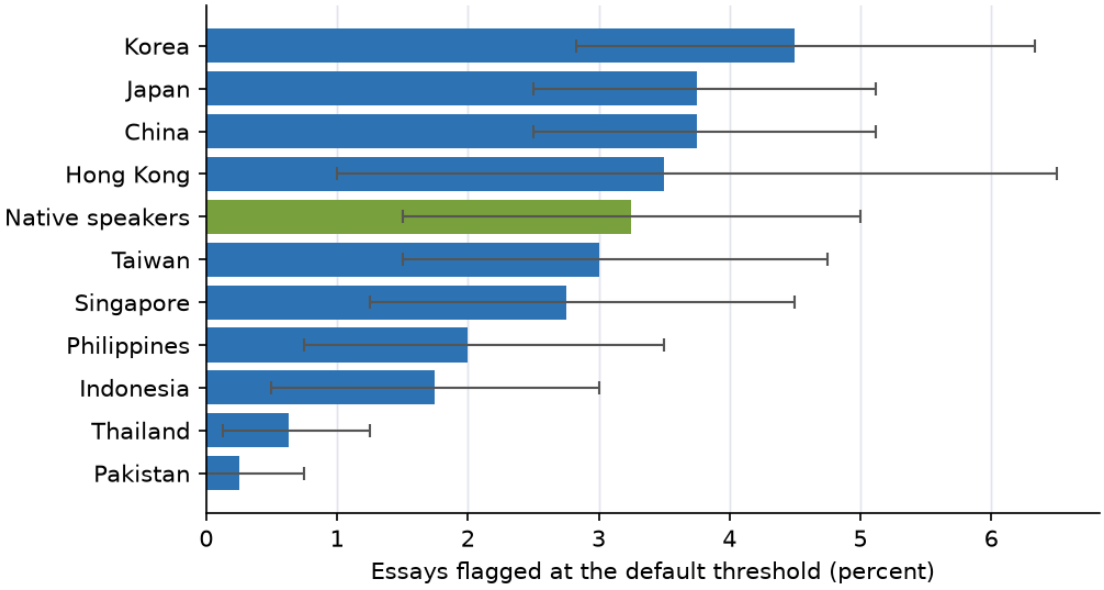

Figure 2. Share of essays flagged at the default threshold by first-language group, with 95 percent student-clustered bootstrap intervals. Native speakers sit inside the learner range.

5. The proficiency gradient runs backwards

If the perplexity mechanism applied here, the least proficient writers would be flagged the most. The opposite occurs. Flag rates increase monotonically with proficiency, from 1.9 percent at A2 (CI 1.0 to 2.8) through 2.4 and 2.8 percent in the two B1 bands to 5.0 percent at B2 and above (CI 3.0 to 7.1). The most proficient learners are flagged two and a half times as often as the least proficient, and more often than the natives.

As learners gain proficiency, their sentences lengthen and even out, their vocabulary diversifies, and their writing style changes toward the polished formal register that the classifier, trained on academic writing and its AI rewrites, associates with an AI-ish writing style. Detection risk grows with mastery of formal English. The gradient survives the length-robustness check. On the concatenated 441-word documents all rates drop sharply, to 0.2 percent for learners pooled and zero for natives, and the B2 band remains the highest learner band.

6. Why learner essays look human to a stylometric perspective

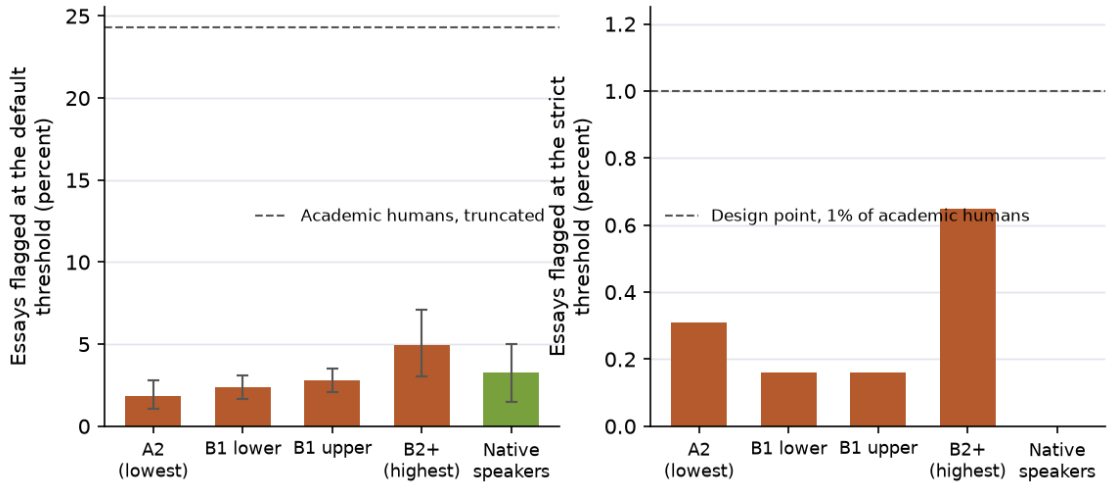

Figure 3. Share of essays flagged by proficiency band, at the default threshold (left) and at the strict threshold designed to flag 1 percent of academic humans (right). The dashed line on the left marks the length-matched academic human baseline.

The logistic regression is linear, so the score gap between any two groups decomposes exactly into per-feature contributions. Decomposing the small learner-native gap shows two opposing forces. The single largest term pushes learners toward an AI verdict, namely, their shorter sentences (0.92 log-odds). Everything else of consequence pushes them toward human. Their lower vocabulary diversity contributes 0.41 log-odds toward human, the two sentence-length dispersion features a further 0.71, led by the learners' higher relative sentence-length variation, and their scarcity of -ly adverbs 0.22. The forces nearly cancel, leaving learners 0.21 log-odds more human-looking than natives on net.

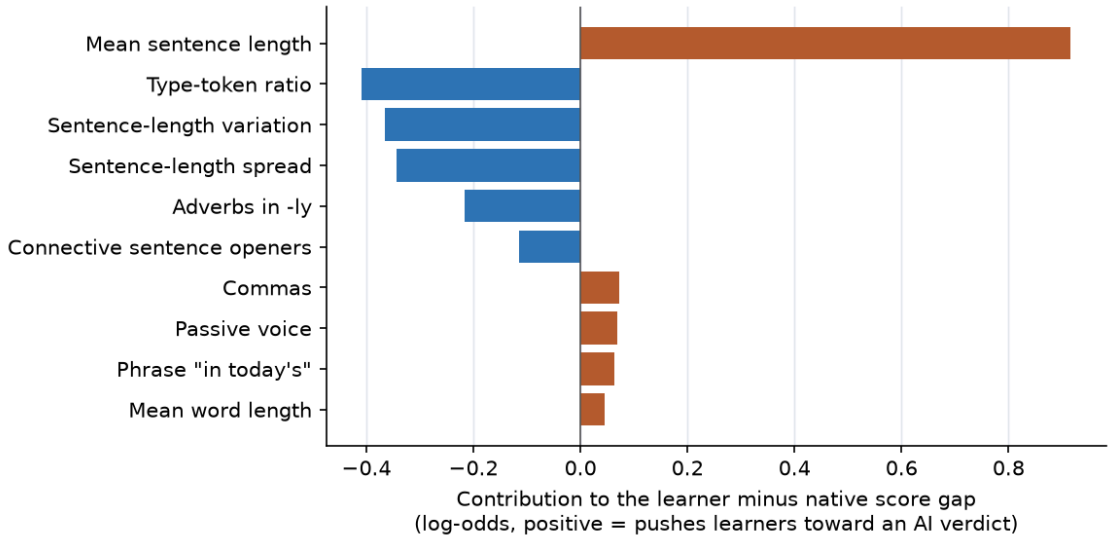

Figure 4. Per-feature contributions to the learner minus native score gap, in log-odds. Positive bars push learner essays toward an AI verdict, negative bars toward human.

The comparison with academic text is more skewed. Both essay groups are farther on the human side of academic writing, impacted by everyday vocabulary, as the scarcity of long words alone contributes 2.1 log-odds toward human and the shorter words 1.4 more, with further contributions from the near-absence of passive voice and the lighter punctuation. The offsetting terms, namely, the essays' more uniform sentence lengths and the high type-token ratio that short texts contain, are smaller in total, which the mean scores confirm.

Under a perplexity perspective, repetitive vocabulary and conventional phrasing read as machine-like predictability, so learner limitations are penalized [1]. Under a stylometric perspective trained on real AI output, the AI signature is diverse vocabulary, uniform sentence rhythm, and polished formality [2, 16], so the same limitations read as human. The features that a non-native writer uses are the features AI models do not produce.

7. The same data, another classifier, the opposite answer

The gradient boosting model, trained on the identical 45 features, the identical texts, and the identical folds, flags 12.5 percent of learner essays (CI 11.6 to 13.5) against 3.8 percent of native essays (CI 1.8 to 6.3), a threefold gap with no overlap in the intervals. Its gradient also points the other way, with the A2 band flagged the most (15.1 percent) and upper B1 the least (10.6 percent). On the same essays where the linear model is fair, the boosted model reproduces the bias of Liang et al. in both direction and structure, though at a lower level than the commercial detectors they tested.

A linear model scores each feature by a global weight, and the learner essays' human-like features outweigh their AI-like ones. A boosted ensemble carves the feature space into regions, and learner essays, which combine short sentences with low diversity in a configuration rare among the academic humans it trained on, are in regions where the training data offers little evidence of humanness. The out-of-distribution writer is penalized not for resembling AI but for resembling nothing in the training data. Its higher in-domain accuracy (AUC 0.968 against 0.950) is of no benefit here. The model that is better on the training distribution is three times more discriminatory off it.

8. Professional human editing increases the linear score

The Edited Essays module pairs 656 learner essays with versions corrected by professional native-speaker editors [15]. Editing moves the logistic regression score up in 61 percent of pairs, with a mean increase of 0.014, and the flag rate at the default threshold nearly doubles, from 2.1 to 4.0 percent. The strict-threshold rate stays near zero in both forms. Grammatical correction and naturalization are movement toward exactly the polished register the linear classifier reads as AI-like.

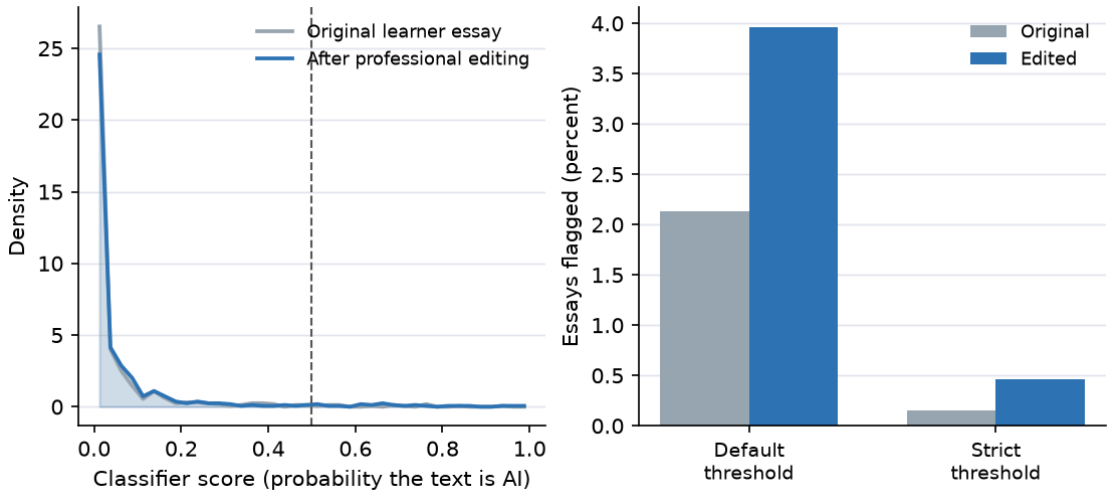

Figure 5. Original learner essays against their professionally edited versions. Left, score distributions. Right, share flagged at the default and strict thresholds.

The gradient boosting model moves in the other direction. Editing lowers its mean score and its flag rate decreases from 12.8 to 9.3 percent, since editing removes the learner-typical irregularities the boosted model penalizes. The same red pen makes an essay more suspicious to one detector and less suspicious to another, a compact illustration that the two models read different things as evidence of AI.

9. Discussion

The question posed by Liang et al. was whether AI detectors are biased against non-native English writers. This study's answer is that the question is underspecified, since the bias is not a property of AI detection as a task. It is a property of the detector. On 5,600 essays where a perplexity mechanism predicts heavy bias against learners, a linear stylometric classifier flags learners no more often than natives, flags the least proficient learners the least, and flags both groups far below its own rate on native-authored academic prose. A boosted model on the same features and data recreates a threefold learner-native gap with the least proficient flagged most.

For deployment, three practical points are worth noting. First, evaluating a detector's fairness requires testing that detector, since neither the feature set nor the training data determines the answer. Second, the fairest configuration here is also the simplest one, and detector builders weighing a few points of in-domain AUC against a threefold difference in group false positive rates should know that trade exists. Third, the writing that this stylometric family of classifiers does over-flag is a formal, polished, native-like writing style, so its fairness problem is a reflection of the published one, involving the strong formal writer rather than the second-language learner, at every proficiency level and in both the native and learner groups.

The proficiency gradient carries a further implication. As second-language writers improve, or as their text is professionally edited, their linear-model detection risk rises toward the formal-writing

baseline. Any process that polishes prose, human or algorithmic, moves text toward the register that stylometric detection reads as AI-ish. The finding of our mixed-authorship (AI + human mixed text) study, that detectors read a document's position on a human-to-AI continuum [10], extends here to a continuum of formality that entirely human writing also traverses.

10. Limitations and conclusion

Four main limitations must be taken into account. First, the essays are short, with a median of 218 words, below the 300-word floor of our prior cohorts, and short inputs both raise false positive rates and widen score dispersion [3]. The length-matched academic baseline addresses the comparison, and the concatenation check shows every conclusion surviving at 441 words, but absolute rates at essay length should be read with the length caveat attached. Second, the corpus covers college students in ten Asian countries and regions writing two argumentative prompts, and other first languages, ages, genres, and registers may place writers differently relative to the training distribution. Third, our classifiers are two members of one transparent stylometric family trained on academic rewriting corpora, and commercial detectors blend other signals, so our rates do not transfer to any specific commercial tool. The structural finding, that the fairness answer flips between two models sharing all inputs, is the transferable part. Fourth, all essays predate modern AI tools, which makes every flag a false positive but leaves detection of actual AI use by learners unmeasured.

A transparent stylometric classifier shows no bias against non-native English writers, flags proficient formal writing the most, and treats learner essays as what they are, namely, strongly human. The published bias against non-native writers is explicit for the AI detectors it was measured on, but it is a consequence of specific detector designs, not a general law of AI detection.

Data availability

The per-essay classifier scores for all 5,600 ICNALE essays, the 656 edited pairs, and the length-matched academic controls (corpus identifiers and scores only, no essay text), the per-group metrics, and all extraction, analysis, and figure code are openly available at https://doi.org/10.5281/zenodo.22061223. The ICNALE corpus itself is distributed by its authors and was used under its terms, which prohibit redistribution of the texts, so it is not included [4, 5]. The training corpora and their feature tables are those of the prior studies [2, 10].

References

[1] 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

[2] 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

[3] 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

[4] Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. Learner Corpus Studies in Asia and the World, 1, 91-118.

[5] Ishikawa, S. (2023). The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English. Routledge. https://doi.org/10.4324/9781003252528

[6] Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., and Khuat, H. Q. (2024). GenAI detection tools, adversarial techniques and implications for inclusivity in higher education. International Journal of Educational Technology in Higher Education, 21, 53. https://doi.org/10.1186/s41239-024-00487-w

[7] 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

[8] Gehrmann, S., Strobelt, H., and Rush, A. (2019). GLTR: Statistical detection and visualization of generated text. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 111-116. https://doi.org/10.18653/v1/P19-3019

[9] Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. Proceedings of the 40th International Conference on Machine Learning, PMLR 202, 24950-24962. https://proceedings.mlr.press/v202/mitchell23a.html

[10] TextPulse Research (2026). The detectability of partially AI-rewritten academic documents from stylometric features. Working paper. https://textpulse.ai/research/ai-partial-rewriting-detection. https://doi.org/10.5281/zenodo.22054615

[11] Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. (2023). Can AI-generated text be reliably detected? arXiv preprint arXiv:2303.11156. https://doi.org/10.48550/arXiv.2303.11156

[12] Dugan, L., Hwang, A., Trhlík, F., Zhu, A., Ludan, J. M., Xu, H., Ippolito, D., and Callison-Burch, C. (2024). RAID: A shared benchmark for robust evaluation of machine-generated text detectors. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 12463-12492. https://doi.org/10.18653/v1/2024.acl-long.674

[13] Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15(4), 474-496. https://doi.org/10.1075/ijcl.15.4.02lu

[14] Crossley, S. A., and McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. Journal of Research in Reading, 35(2), 115-135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

[15] Ishikawa, S. (2018). The ICNALE Edited Essays: A dataset for analysis of L2 English learner essays based on a new integrative viewpoint. English Corpus Studies, 25, 117-130.

[16] 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