

Fixed by AI, Flagged as AI: A Matched Audit of AI-Text Detectors on 12,132 Student Essays

Zakhar Sukhanov

Independent researcher · zexa1993@outlook.com · [LinkedIn](#) · [GitHub](#)

Preprint · DOI [10.5281/zenodo.23400437](https://doi.org/10.5281/zenodo.23400437) · 15 October 2026

Abstract.

A 2023 Stanford study (Liang et al., 2023) reported that AI-text detectors flag most essays by non-native English writers as AI-generated: seven detectors labeled 61% of 91 TOEFL essays as AI-written while almost never flagging essays by US eighth graders. But the two sets differ in more than the writer’s first language: the TOEFL essays are short test answers on different topics. As a high school student applying to US universities in my second language, I wanted to know whether essays like mine would be judged unfairly, so I repeated the test with a fairer comparison: 12,132 student essays, including 8,244 in which English learners and native speakers in US schools answered the same prompts, 2,989 AI-written essays on those prompts, and 9 open detectors released between 2019 and 2025. On the Stanford TOEFL essays the result holds: trained detectors flag 43%–97% of them. On matched essays it does not: 7 of 9 detectors flag English learners significantly less often than native speakers, and none flags them more often. What does trigger false flags is the text itself. Short essays are flagged more often, and so are essays touched by a language model: when an AI tool corrected only the grammar of students’ essays, leaving most words unchanged, every detector flagged more of them, up to 57 times more for English learners’ essays (Mozilla’s Fakespot detector: 0.5% → 28.5%). Fixing spelling with a dictionary changed almost nothing, and the rewording that the Stanford study found to reduce false flags now increases them for every detector. In these data, detectors do not penalize non-native writers; they penalize text that is short, formulaic, or polished by AI. A flag on a carefully edited essay, such as a college application, says little about who actually wrote it.

1 Introduction

Schools and universities now run student writing through AI-text detectors, tools that estimate whether a passage was written by a large language model (LLM). For the student involved, a false positive is not a statistic: it can mean an academic-integrity case, a lower grade or, in admissions, an essay read with suspicion. The best-known evidence on who bears these errors is a 2023 Stanford study by Liang et al. (2023). They ran seven detectors on 91 TOEFL essays by non-native writers and 88 essays by US eighth graders. The detectors flagged 61% of the TOEFL essays as AI-written on average and almost none of the US essays. The study is now the standard citation for the claim that detectors are biased against non-native English writers, and detector companies answer it directly: Pangram reports no false positives on the same TOEFL set (Emi & Spero, 2024) and GPTZero reports an “ESL de-biasing” step (GPTZero, 2023).

The comparison behind that claim changes many things at once. The TOEFL essays are short answers written under time pressure on test prompts and collected from a test-preparation forum; the US essays are longer school assignments on other topics. Writer background, length, genre and source all differ, so the data cannot say which of them the detectors respond to. The companies’ replies are tested on the same 91 essays.

The question matters to me directly. I am a high school student from Russia applying to US universities this year, and I am writing my application essays in English, my second language. If detectors really treat non-native writing as machine-written, essays like mine are at risk; if they respond to something else, students and admissions readers should know what that is. So I asked the question again with a design that holds topic and setting fixed. I use the PERSUADE 2.0 corpus (Crossley et al., 2024), in which US students in grades 6–

12 who are and are not classified as English language learners (ELL) wrote on the same 15 prompts, and the ELLIPSE corpus (Crossley et al., 2023), in which ELL students’ essays carry expert ratings of English proficiency. I score 12,293 human essays and 2,989 LLM-written essays on the same prompts with 9 open detectors released between 2019 and 2025, including the classifier behind Mozilla’s Fakespot browser add-on (Fakespot AI Labs, 2025) and two recent zero-shot methods (Bao et al., 2024; Hans et al., 2024). Then I change the human essays directly: I cut them to TOEFL length, and I let an LLM fix only their spelling and grammar, the way a student might with a writing assistant.

The findings are:

1. **The original result replicates on the original data.** At their default settings, the six trained classifiers flag 43%–97% of Liang et al.’s TOEFL essays (mean 64%).
2. **It does not hold for English learners writing on the same prompts as native peers.** With prompt, grade and length held fixed, 7 of 9 detectors flag ELL essays significantly *less* often than native-speaker essays, and none flags them more often.
3. **What the detectors respond to is the text.** Short length raises false positives for nearly every detector. Published “model” college-admission essays, the most heavily edited human texts in the study, are flagged at up to 83%. For 8 of 9 detectors, English learners’ essays look more AI-like the higher their rated proficiency, and three trained classifiers flag native speakers’ essays more often the higher human raters scored them; the newest classifier shows the opposite trend.
4. **A language model’s light touch is enough.** Letting an LLM fix only spelling and grammar, which leaves most words in place, multiplies the share of human essays flagged for all 9 detectors; correcting spelling with a dictionary does not. The rewording that Liang et al. found to reduce false positives now backfires.

Figure 1 shows the last finding on one real essay.

Although some might argue that schools make students stressed, ~~however,~~ schools provide entertaining trips for students, as mentioned ~~before, schools~~ **before.** **Schools** is a place where students make friendships. A trip with friends helps to overcome stress. ~~Not to add,~~ **Additionally,** schools also offer lots of programs and counseling to help out stressed students to overcome their stress.

Detector	As written	After LLM grammar fix
Desklib (2025)	0.7% (human)	62.7% (AI)
Fakespot / Mozilla (2025)	0.1% (human)	86.5% (AI)
OpenAI RoBERTa (2019)	0.0% (human)	100.0% (AI)

Figure 1: Second paragraph of a 412-word essay by a grade-12 English learner (PERSUADE 2.0, prompt “Distance learning”), with the edits an LLM made when asked to fix only grammar: deletions struck through, insertions highlighted. The table gives each detector’s probability that the whole essay is AI-written, and its verdict at the threshold that flags 5% of native essays. A dictionary spellcheck found nothing to change in this essay.

The practical message for students, teachers and admissions readers is that a false flag says more about the text than about who wrote it, and that a student who lets an AI tool fix the grammar of an essay they wrote can get it flagged.

2 Related work

Detectors. Trained detectors fine-tune an encoder on human and machine text: the GPT-2 output detector (Solaiman et al., 2019), models trained on ChatGPT answers (Guo et al., 2023), adversarially trained RADAR (Hu et al., 2023), and recent classifiers trained on the RAID benchmark (Desklib, 2025; Dugan et al., 2024; Fakespot AI Labs, 2025; SuperAnnotate, 2024). Zero-shot detectors instead read the text with a language model and test whether it is unusually predictable: DetectGPT (Mitchell et al., 2023) and Fast-DetectGPT (Bao et al.,

2024) compare the text’s likelihood to that of alternative tokens, and Binoculars (Hans et al., 2024) normalizes perplexity by the cross-perplexity between two related models.

Fairness. Liang et al. (2023) linked the bias they found to perplexity: TOEFL essays with simpler vocabulary are more predictable, and rewriting them with richer word choice lowered the false-positive rate. Vendors’ reports of fixed bias evaluate their own products (Emi & Spero, 2024; GPTZero, 2023). An independent comparison of commercial detectors on human text from several genres found very low error rates for the best tool but did not study non-native writing (Jabarian & Imas, 2025). Closest to this paper, Park et al. (2026) compare scientific manuscripts by non-native authors before and after professional English editing and find that editing moves detector scores, in different directions for different detectors. I study student essays rather than research papers, compare learners and native speakers on identical prompts, relate flags to human ratings of writing quality, and decompose the original TOEFL result into length and text-type effects.

3 Data

Native and ELL essays (PERSUADE 2.0). The corpus contains about 25,000 argumentative essays by US students in grades 6–12 on 15 prompts, collected before the release of ChatGPT, with each writer’s ELL status, grade and a holistic score from 1 to 6 assigned by trained raters (Crossley et al., 2024). I keep all 2,244 essays by ELL students and a random sample of 6,000 by non-ELL students (hereafter “native”), each with at least 100 words. The native sample is split in half at random: 3,000 essays are used only to set detector thresholds, and the rest are used for evaluation.

ELL proficiency (ELLIPSE). 3,888 essays by ELL students in grades 8–12, each rated by trained raters from 1 to 5 for overall proficiency and for cohesion, syntax, vocabulary, phraseology, grammar and conventions (Crossley et al., 2023). Some ELLIPSE essays also appear in PERSUADE; the analyses keep the two corpora separate.

Liang et al.’s essays. The 91 TOEFL essays (median 104 words) and the 70 “authentic” US college-admission essays released with Liang et al. (2023). The admission essays are published examples from two admissions-advice websites (“college essay examples that worked”), so they are polished, often professionally edited texts rather than a random sample of applicants’ writing.

AI essays. 2,989 essays written by 15 LLM configurations (including ChatGPT, GPT-4, Claude, PaLM, Llama 2 and Mistral variants) on the same PERSUADE prompts, sampled up to 200 per source from the DAIGT v2 training set of a 2023 Kaggle competition (Kaggle, 2023). They are used to check that each detector works on this kind of text and to report its true-positive rate.

4 Detectors and thresholds

I run 9 detectors locally (Tables 1 and 2). Six are trained classifiers that output a probability that the text is AI-written: the 2019 GPT-2 output detector (Solaiman et al., 2019), a RoBERTa model trained on the HC3 corpus (Guo et al., 2023), RADAR (Hu et al., 2023), SuperAnnotate’s RoBERTa-large detector (SuperAnnotate, 2024), the RoBERTa detector released with Mozilla’s Fakespot add-on (Fakespot AI Labs, 2025), and Desklib’s DeBERTa-v3-large detector, which led the RAID leaderboard when released (Desklib, 2025). Three are zero-shot scores computed from open language models: log-perplexity under Qwen2.5-1.5B (Qwen Team, 2024); Binoculars (Hans et al., 2024) with Qwen2.5-1.5B as observer and Qwen2.5-1.5B-Instruct as performer; and analytic Fast-DetectGPT (Bao et al., 2024) with Qwen2.5-1.5B as both sampling and scoring model. Texts are truncated to 512 tokens. The zero-shot detectors use smaller models than their original papers because of hardware limits, which lowers their accuracy (Table 2).

A detector’s false-positive rate depends on where its threshold is set, so I report three settings. **Default** uses a trained classifier’s own decision rule, $P(\text{AI}) > 0.5$, which is how such a classifier is used out of the box. **1%** and **5% FPR** set the threshold so that 1% or 5% of the native calibration essays are flagged, which puts all

ten detectors on the same footing; rates on the held-out native essays then measure how well that calibration transfers.

I summarize group differences in three ways: the share of each group’s essays above the threshold, with Wilson 95% intervals; the ratio of the ELL rate to the native rate; and a logistic regression of being flagged on ELL status with fixed effects for prompt and controls for grade and log length, with and without the essay’s holistic score.

5 Results

5.1 *The original result replicates on the original essays*

Detector	AUROC	Native	ELL	TOEFL	College	AI caught
OpenAI RoBERTa (2019)	0.90	4.0%	1.9%	58%	1%	69%
HC3 RoBERTa (2023)	0.78	14.9%	12.9%	43%	0%	67%
RADAR (2023)	0.87	30.4%	32.0%	97%	4%	85%
SuperAnnotate (2024)	0.98	29.3%	12.7%	64%	83%	100%
Fakespot/Mozilla (2025)	1.00	8.7%	1.7%	69%	29%	100%
Desklib (2025)	1.00	5.4%	5.7%	53%	21%	100%

Table 1: Trained classifiers at their default threshold ($P(\text{AI}) > 0.5$). AUROC: how well the detector separates native PERSUADE essays from AI essays on the same prompts (1 = perfectly). Native, ELL, TOEFL, College: share of human essays flagged as AI. AI caught: share of AI essays flagged.

All six trained classifiers separate human from AI essays on these prompts better than chance, and the three most recent do so almost perfectly (Table 1). On Liang et al.’s TOEFL essays they reproduce the original finding: at the default threshold they flag 43%–97% of them, against a mean of 61% for the seven commercial detectors of 2023. The same tools flag published college-admission essays at very different rates, from 0% to 83%.

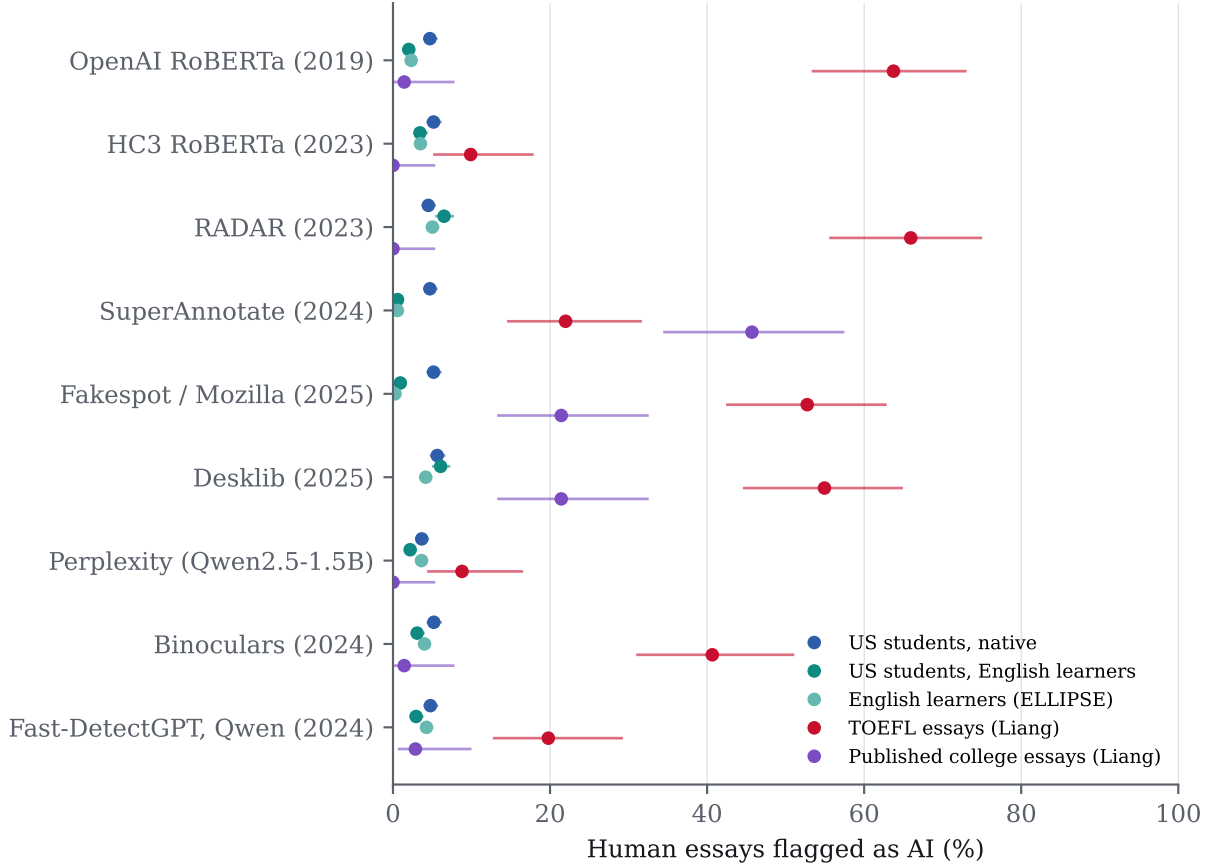


Figure 2: Share of human essays flagged as AI by each detector when its threshold is set to flag 5% of native calibration essays. Lines are 95% intervals.

5.2 English learners writing on the same prompts are flagged less, not more

With every detector calibrated to flag 5% of native essays, the held-out native essays are flagged at close to 5% and ELL essays on the same prompts at lower rates for 7 of 9 detectors (Figure 2). The regression holds prompt, grade and length fixed: the odds ratio for ELL status is significantly below one for 7 of 9 detectors and significantly above one for none (Table 2). The ELLIPSE essays, all by English learners, are flagged at or near the native rate. Without any threshold, the probability that a random ELL essay scores as more AI-like than a random native essay on the same prompt is below one half for 7 of 9 detectors.

Detector	Native	ELL	ELLIPSE	TOEFL	AI caught	Odds ratio, ELL (95% CI)
OpenAI RoBERTa (2019)	4.7%	2.0%	2.3%	64%	70%	0.47 (0.32–0.70)
HC3 RoBERTa (2023)	5.2%	3.4%	3.5%	10%	55%	0.42 (0.30–0.59)
RADAR (2023)	4.5%	6.5%	5.0%	66%	60%	1.02 (0.77–1.36)
SuperAnnotate (2024)	4.7%	0.6%	0.6%	22%	98%	0.09 (0.05–0.17)
Fakespot/Mozilla (2025)	5.2%	0.9%	0.3%	53%	100%	0.12 (0.08–0.21)
Desklib (2025)	5.6%	6.1%	4.2%	55%	100%	0.82 (0.63–1.08)
Log-perplexity	3.7%	2.2%	3.6%	9%	81%	0.57 (0.38–0.86)
Binoculars	5.2%	3.1%	4.0%	41%	57%	0.56 (0.40–0.79)
Fast-DetectGPT (Qwen)	4.8%	2.9%	4.3%	20%	51%	0.59 (0.41–0.85)

Table 2: All detectors with the threshold set to flag 5% of native calibration essays. Odds ratio: being flagged, ELL vs native, from a logistic regression with prompt fixed effects, grade and log word count (PERSUADE held-out essays).

5.3 Writing quality matters, but not in the same way for every detector

If the detectors responded to non-native writing as such, the weakest English learners should look most AI-like. For most detectors the opposite holds. In ELLIPSE, the score of 8 of 9 detectors is positively rank-correlated with the writer’s rated overall proficiency (Spearman’s ρ from 0.08 to 0.52); for the Desklib classifier it is flat ($\rho = -0.08$). Among native speakers in PERSUADE, the HC3, SuperAnnotate and Fakespot classifiers flag essays more often the higher their holistic score: Fakespot flags 1.7% of essays scored 1 and 17.2% of essays scored 6, and SuperAnnotate 0.0% and 26.3% (Figure 3). Desklib, the newest and most accurate classifier, shows the reverse: it flags 13.9% of native essays scored 1 and 2.0% of those scored 6. The older classifiers and the zero-shot scores show no clear trend. Writing quality is therefore not a universal driver of false positives; which human texts a detector confuses with AI depends on what human text it was trained on. Controlling for the holistic score in the regression moves the ELL odds ratios toward one for most detectors but does not remove the gap.

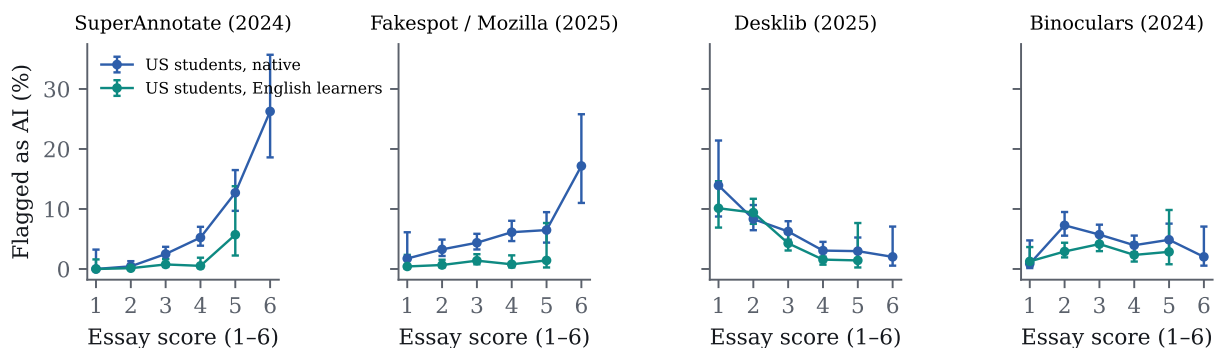


Figure 3: Share of PERSUADE essays flagged (5%-FPR threshold) by holistic score assigned by trained human raters, for four detectors. Points with fewer than 30 essays are omitted.

5.4 Length explains part of the TOEFL result

The TOEFL essays are short (median 104 words, against 375 for native PERSUADE essays). To separate length from writer group, I cut 1,000 native and 1,000 ELL essays to TOEFL-like length: each keeps its first k words, with k drawn from the TOEFL length distribution, and is trimmed back to the last full sentence. Shortening raises false positives for both groups with seven of the nine detectors; only the log-perplexity and Fast-DetectGPT scores do not rise (Table 3). For the Fakespot detector, for example, the share of native essays flagged rises from 5.2% to 17.9%, and for RADAR it jumps from 4.3% to 78.1%. At the same length, ELL essays are flagged no more often than native ones. For most detectors the shortened essays are still flagged less often than the TOEFL essays (Fakespot: 17.9% against 52.8%). Length therefore accounts for part of the original gap, and the rest belongs to the TOEFL essays themselves: they are timed answers built from stock phrases, and they are also unusually clean (a dictionary spellchecker finds nothing to correct in them; see below).

5.5 An LLM grammar fix turns human essays into “AI” essays

Students are told to proofread, and many now do it with writing tools that use language models. I took 200 ELL essays, 200 native essays on the same prompts and the 91 TOEFL essays, and asked an open model (Qwen2.5-7B-Instruct, run locally) to “correct the spelling, grammar and punctuation mistakes ... Do not change word choice, sentence structure, tone or meaning, and do not add or remove content.” The edits are light: the median share of words kept in order is 0.81 for ELL essays, 0.88 for native essays and 0.94 for TOEFL essays. As a control without any language model, I corrected only misspelled lowercase words with a dictionary spellchecker (pyspellchecker), which changes about 2% of the words in ELL essays.

The LLM correction raises false positives for every detector and both groups (Table 3). With every threshold set to flag 5% of native essays, the share of ELL essays flagged rises by a factor of 3 to 57 across the 9 detectors, and of native essays by 3 to 17. The Fakespot detector flags 0.5% of the ELL essays before correction and 28.5%

after it; Desklib goes from 6.5% to 70.0%. The dictionary correction raises rates by more than two points for only 3 detectors on ELL essays and for none on native essays. So the detectors are not reacting to the absence of spelling errors as such: they react to text that has passed through a language model, even when it changed only a few words. For the Fakespot and SuperAnnotate classifiers, LLM-corrected essays in which fewer than 5% of words changed are flagged at least as often as more heavily corrected ones ([grammar_by_edit.csv](#)), although these lightly edited essays also started out cleaner.

Detector	Cut to TOEFL length		Dictionary spellcheck		LLM grammar fix	
	Native	ELL	Native	ELL	Native	ELL
OpenAI '19	4.5% → 19.1%	2.3% → 11.7%	4.5% → 5.5%	2.5% → 5.5%	4.5% → 48.0%	2.5% → 64.0%
HC3 '23	5.6% → 13.4%	3.0% → 9.3%	3.5% → 3.5%	2.0% → 6.0%	3.5% → 21.5%	2.0% → 45.0%
RADAR '23	4.3% → 78.1%	6.6% → 80.8%	2.0% → 4.0%	7.5% → 10.5%	2.0% → 13.5%	7.5% → 24.0%
SuperAnn. '24	4.4% → 9.0%	0.4% → 0.8%	1.5% → 1.5%	0.5% → 0.5%	1.5% → 18.5%	0.5% → 10.5%
Fakespot '25	5.2% → 17.9%	1.0% → 5.2%	4.5% → 4.5%	0.5% → 0.5%	4.5% → 28.5%	0.5% → 28.5%
Desklib '25	4.9% → 45.8%	6.2% → 46.7%	3.0% → 3.5%	6.5% → 5.0%	3.0% → 51.5%	6.5% → 70.0%
Log-ppl	4.0% → 2.0%	1.6% → 0.4%	1.5% → 1.5%	2.5% → 2.5%	1.5% → 16.0%	2.5% → 15.5%
Binoculars	6.0% → 14.6%	3.0% → 8.4%	6.0% → 6.0%	4.5% → 5.5%	6.0% → 18.5%	4.5% → 24.5%
Fast-Detect	5.7% → 6.9%	2.5% → 3.0%	6.0% → 6.0%	4.5% → 4.5%	6.0% → 19.0%	4.5% → 20.5%

Table 3: Share of the same human essays flagged before → after each edit, with thresholds set to flag 5% of native calibration essays. Length: 1,000 essays per group. Spellcheck and LLM fix: the same 200 essays per group.

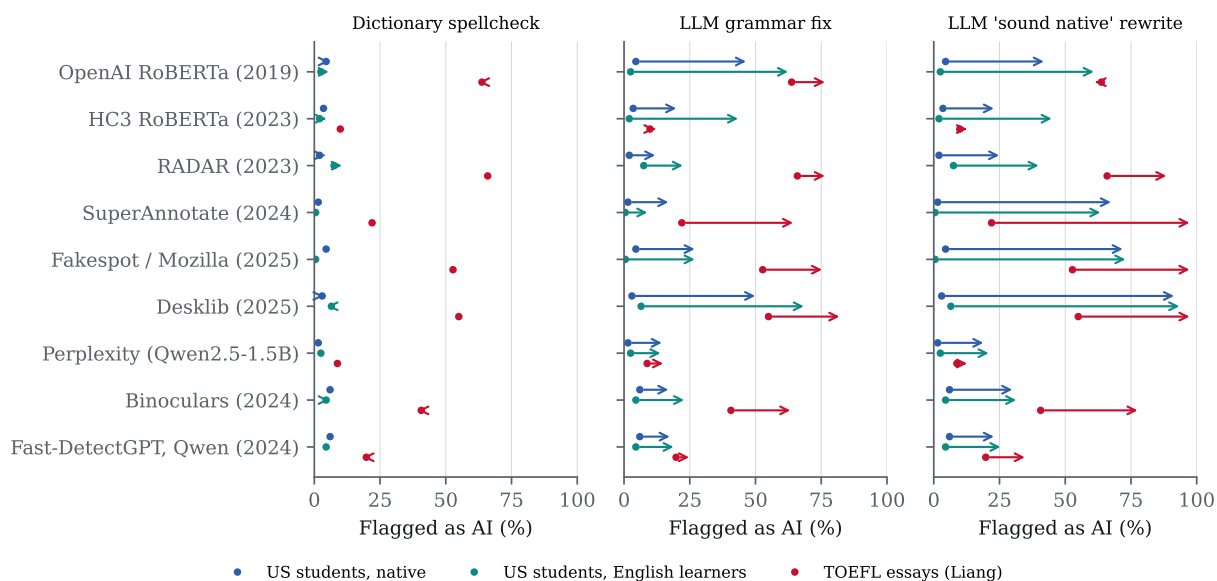


Figure 4: The same human essays before (dot) and after (arrow tip) each edit, with thresholds set to flag 5% of native calibration essays: 200 ELL essays, 200 native essays and the 91 TOEFL essays.

5.6 The 2023 remedy now backfires

Liang et al. (2023) showed that asking ChatGPT to “enhance the word choices to sound more like that of a native speaker” made their detectors flag far fewer TOEFL essays, and concluded that such prompting can mitigate the bias. I applied the same prompt with the local model. Unlike the grammar fix, it rewrites heavily (median share of words kept in order 0.54 for ELL essays). With today’s detectors it has the opposite effect: the share of ELL essays flagged rises for all 9 detectors, by a factor of 6 to 149, and Desklib flags 95% of the rewritten essays. Detectors trained on modern LLM output recognize the rewriting itself.

6 Discussion

Who is at risk. I started this project expecting to confirm that detectors are unfair to writers like me. In these data they are not, at least once topic and setting are held fixed. English learners' essays carry spelling mistakes, unusual word choices and uneven sentences, and these make a text less predictable to a language model and less like the machine-written examples a classifier learned from. The human texts that detectors mistake for AI are of three other kinds: short texts; formulaic ones, such as timed-test answers built from stock phrases and heavily edited showcase essays; and texts that a language model has touched, even lightly. Exactly which polished texts a detector flags differs from one detector to another, as the opposite quality trends of Fakespot and Desklab show. So the bias that Liang et al. (2023) found is real for the kind of text they tested, but it is better described as a bias against short, formulaic or machine-smoothed writing than against a group of writers.

Proofreading with AI. The grammar experiment points to a risk that nobody warns students about. Many writing tools now offer AI rewriting right next to ordinary spellchecking, and the two look like the same kind of help. They are not, as far as detectors are concerned. A student who accepts an AI tool's grammar fixes on an essay they wrote themselves can turn it from an essay few detectors flag into one that many flag, while a dictionary spellcheck barely changes the verdict. Whether such an essay should count as "AI-written" is a question for schools, but a detector cannot tell it apart from one a model wrote, and the student may not know that the choice of tool matters.

Why this matters for admissions. Application essays are probably the most edited texts a student ever writes. They go through many drafts, teachers, counselors and grammar checkers. In this study, published examples of successful essays, the kind applicants read and imitate, were flagged by several detectors at rates no admissions office would accept. A detector that is nearly perfect at telling ordinary school essays from AI essays can still be wrong about exactly these texts, because they are unlike the human writing it was trained on. For me as an applicant, the practical lesson is simple: proofread with a plain spellchecker, keep my drafts, and do not let an AI tool "fix" an essay I wrote. For admissions readers, a flag on an application essay should not be read as evidence about its author without other information.

What vendors' claims can and cannot show. Reports that a detector has no false positives on Liang et al.'s 91 TOEFL essays (Emi & Spero, 2024; GPTZero, 2023) test a single small set that was published three years ago. The results here suggest checking detectors on human text that varies in quality and amount of editing, with groups matched on topic, rather than on one benchmark of non-native writing.

Recommendations. For schools and universities: do not act on a detector flag alone; calibrate any detector on human writing from the same setting and report its false-positive rate there; treat short, formulaic and AI-proofread writing as known sources of false positives; and tell students that AI grammar correction can get an essay they wrote flagged. For detector developers: include edited and high-scoring human writing, and short texts, among the human training and test data.

7 Limitations

The study uses only open detectors. Commercial detectors used by schools, such as Turnitin, GPTZero and Pangram, may behave differently; testing them on the same matched essays is the most important next step. The zero-shot detectors use smaller language models than their original papers, and all texts are truncated to 512 tokens. The ELL label in PERSUADE comes from US school records and covers students who are learning English inside the US school system; international applicants writing in English as a foreign language may differ. Liang et al.'s college essays are published examples, not a random sample of applications, and they may appear in the training data of the language models behind some detectors. The PERSUADE and ELLIPSE essays have been public since 2022 and may also have been seen in training, which would lower their apparent AI-likeness; this affects ELL and native essays from the same corpus alike. The AI essays come from 2023 models. Grammar correction was done by a local open model, not by a commercial grammar checker, and is a proxy for that use.

8 Statement on AI assistance

This project was carried out with the help of Claude (Anthropic), an AI assistant. I am responsible for its content.

9 Data and code availability

All code is available at github.com/Dek0rta/flagged. The essays are not redistributed; a script downloads them from their public sources (PERSUADE 2.0 and ELLIPSE under CC BY-NC-SA 4.0, Liang et al.’s essays under the MIT license, and the DAIGT v2 essays). Per-essay detector scores are released with the code, keyed by the corpora’s essay identifiers.

References

- Bao, G., Zhao, Y., Teng, Z., Yang, L., & Zhang, Y. (2024,). Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *International Conference on Learning Representations (ICLR)*.
- Crossley, S. A., Baffour, P., Tian, Y., Franklin, A., Benner, M., & Boser, U. (2024). A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Assessing Writing*, 61.
- Crossley, S. A., Tian, Y., Baffour, P., Franklin, A., Kim, Y., Morris, W., Benner, M., Picou, A., & Boser, U. (2023). Measuring second language proficiency using the English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) corpus. *International Journal of Learner Corpus Research*, 9(2), 248–269.
- Desklib. (2025,). *desklib/ai-text-detector-v1.01 (model card)*. <https://huggingface.co/desklib/ai-text-detector-v1.01>
- Dugan, L., Hwang, A., Trhlik, F., Ludan, J. M., Zhu, A., Xu, H., Ippolito, D., & 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 (ACL)*.
- Emi, B., & Spero, M. (2024). *Technical report on the Pangram AI-generated text classifier*.
- Fakespot AI Labs. (2025,). *ApolloDFT: AI text detector (fakespot-ai/roberta-base-ai-text-detection-v1)*. <https://github.com/FakespotAILabs/ApolloDFT>
- GPTZero. (2023,). *5 steps towards responsible AI detection*. <https://gptzero.me/news/5-steps-towards-responsible-ai-detection>
- Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). *How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection*.
- Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J., & Goldstein, T. (2024,). Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text. *Proceedings of the 41st International Conference on Machine Learning (ICML)*.
- Hu, X., Chen, P.-Y., & Ho, T.-Y. (2023,). RADAR: Robust AI-text detection via adversarial learning. *Advances in Neural Information Processing Systems (Neurips)*.
- Jabarian, B., & Imas, A. (2025). *Artificial writing and automated detection* (No. NBER Working Paper 34223). National Bureau of Economic Research.
- Kaggle. (2023,). *LLM - Detect AI Generated Text (Kaggle competition) and the DAIGT v2 training set*. <https://www.kaggle.com/competitions/llm-detect-ai-generated-text>
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7). <https://doi.org/10.1016/j.patter.2023.100779>

- Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023,). DetectGPT: Zero-shot machine-generated text detection using probability curvature. *Proceedings of the 40th International Conference on Machine Learning (ICML)*.
- Park, H., Jeong, G., & Kim, B. (2026). *Style as a confound: False positives in AI detection of non-native academic writing*.
- Qwen Team. (2024). *Qwen2.5 technical report*.
- Solaiman, I., Brundage, M., Clark, J., Askill, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., & Wang, J. (2019). *Release strategies and the social impacts of language models*.
- SuperAnnotate. (2024,). *SuperAnnotate/ai-detector (model card)*. <https://huggingface.co/SuperAnnotate/ai-detector>