Do AI Detectors Agree?

Do AI Detectors Agree? An Inter-Rater Reliability Study of Commercial AI Text Detectors on Academic Writing

TextPulse Research · Working Paper

1. Abstract

AI text detectors are currently embedded in plagiarism tools, admissions workflows, and editorial pipelines, where a single tool's verdict can carry disciplinary consequences in the long run. A growing body of literature evaluates how accurately individual detectors separate human from AI-generated text, but research on whether detectors agree with one another is lacking. Therefore, this study employs nine commercial detectors, namely, Turnitin, GPTZero, Originality.ai, Pangram, Copyleaks, ZeroGPT, Winston, Sapling, and QuillBot, as independent raters of the same texts, and applies standard inter-rater reliability methods to their final scores. We assembled 90 research-style passages in three conditions: (1) human-written text published before 2022; (2) text generated by five current model families; and (3) hybrid text pieces splicing the two. Every text was scored by each of the detectors. In terms of overall agreement, Krippendorff's alpha was 0.71, Fleiss' kappa 0.78, and mean pairwise Cohen's kappa 0.73. In contrast, the average demonstrates disagreement. On purely human and purely AI texts, detectors agreed on 96 to 97% of pairwise verdicts. However, on hybrid texts, agreement was generally by chance level (mean pairwise kappa 0.02). The majority (28 of 30) of the hybrid texts received both a "human" verdict and an "AI" verdict from different tools, and the median per-text score range was 100 points on a 100-point scale. Across the full corpus, 41% of texts received contradictory judgement. Agreement was equally poor at every alternative decision threshold, which indicates that detectors disagree about the texts themselves, not about cutoffs. On mixed human-AI writing, the classification a text receives relies more on the choice of detector than on the text itself. We discuss the reason single-detector verdicts are unsafe as evidence in academic settings, and make the full score matrix and corpus publicly available for reanalysis.

2. Introduction

Since the release of ChatGPT at the end of 2022, institutions that assess writing have faced a vital question: can they tell whether a given text was actually written by a person? A market of

AI text detectors has evolved around this question. Turnitin reports having screened millions of student text submissions for AI involvement. GPTZero, Originality.ai, Copyleaks, ZeroGPT, and a rapidly growing list of competitors assist universities, publishers, hiring teams, and search engine SEO content businesses. In many of these settings, the detector's output is not advisory. The final judgement they offer in the form of a "human score" or "AI score" can lead to a case of academic misconduct.

The research response to these tools has concentrated mostly on accuracy. Studies benchmark detectors against texts of known origin, and then report how often each tool is correct (Weber-Wulff et al., 2023; Elkhatat et al., 2023; Walters, 2023; Dugan et al., 2024; Jabarian & Imas, 2025). This work produced two consistent findings. First, accuracy varies widely across these tools, from high performance for the strongest commercial detectors on raw AI text (Jabarian & Imas, 2025), to chance based performance for weaker tools and for text that has been mixed with human writing, paraphrased, or translated to another language (Weber-Wulff et al., 2023; Hadra et al., 2026). Second, errors are not evenly distributed. Detectors excessively flag writing by non-native English speakers (Liang et al., 2023), and their performance decreases abruptly on hybrid human-AI texts, which makes up the majority of real student work (Van Vlasselaer et al., 2026; Hadra et al., 2026).

Accuracy however is only one part of what makes a measurement tool reliable. The other part is reliability. Does the instrument itself produce consistent results across equivalent measurements? In every field that uses classification judgments with consequences, such as medical imaging and content analysis, consistency between independent raters is treated as a precondition for taking any single rater's judgment seriously, and this is measured using standard statistics such as Cohen's kappa, Fleiss' kappa, and Krippendorff's alpha (Cohen, 1960; Fleiss, 1971; Krippendorff, 2019). AI detectors are considered 'raters' in practice. Different institutions license different tools, so the same essay that passes one university's detector can be flagged by the tool at another university. If detectors systematically disagree, then the outcome of an integrity case depends less on what the student wrote than on which rater the institution uses.

To the best of our knowledge, no published study has measured this. Comparative evaluations run multiple detectors across the same corpus, which produces the data needed for agreement analysis, but they only report each tool's accuracy against ground truth. None applies inter-rater reliability statistics to the detectors themselves (Weber-Wulff et al., 2023; Dugan et al., 2024; Van Vlasselaer et al., 2026; Hadra et al., 2026). This empirical gap has practical value. Institutions currently have no evidence base for questions they face daily. If two tools disagree, is that unusual? Should a second detector's "human" verdict be considered?

This study aims to address this gap. We employ nine widely used commercial detectors as independent raters of a common corpus of 90 research-style texts spanning (1) pre-LLM, unmistakably human writing, (2) output from five prominent model families, and (3) hybrid

human-AI texts; and we apply inter-rater reliability metrics to their scores and verdicts. The two research questions are formulated as follows:

RQ1. How strongly do commercial AI detectors agree with one another when classifying identical academic texts, overall and within each authorship condition?

RQ2. How often do detectors return contradicting judgements (scores) on the same text, and how large are per-text score ranges?

Along with the agreement analysis we report conventional accuracy metrics, and this allows a direct comparison with prior benchmarks. We make the complete detector-by-text score matrix and corpus with per-item provenance publicly available for independent re-analysis.

3. Literature review

3.1 Detection approaches and commercial tools

The automated detection of machine-generated text predates modern chatbots, but the current tool market formed after ChatGPT's release during the end of 2022. Detectors fall into three primary approaches: (1) statistical approaches that score text against the probability distributions of language models, using measures such as perplexity and the variability of sentence-level surprisal, or "burstiness"; (2) supervised classifiers trained on labelled human and AI corpora; and (3) watermarking schemes that require cooperation from the text generator and are absent from deployed commercial products (Sadasivan et al., 2023; Dugan et al., 2024). Commercial vendors do not disclose their architectures, training data, or decision thresholds. Tools return outputs on different scales, such as a probability that the text is AI-generated, a percentage of the text estimated to be AI-written, or a discrete, categorical label. Each vendor chooses its own cutoff for converting a score into a final judgement. This heterogeneity is one motivation for treating the tools as "black-box" raters, as the present study does. Regardless of the architecture or scoring approach, a final verdict is generated to score an input text, and this score can be compared across detectors.

3.2 Accuracy evaluations and their limits

Independent evaluations of detector accuracy started within months of ChatGPT's release. Weber-Wulff et al. (2023) tested 14 tools, including Turnitin and PlagiarismCheck, and concluded that available detectors were "neither accurate nor reliable," with a bias toward labelling text as "human-written". Elkhatat et al. (2023) found that the same tools that performed adequately on GPT-3.5 output deteriorated on GPT-4, and observed false positives for classification on human control texts. Walters (2023) and Chaka (2023) reported similar heterogeneity across tools and text types. At broader scale, the RAID benchmark (Dugan et al., 2024) evaluated 12 detectors over six million generations across 11 generator models, 8

domains, and 11 adversarial attacks, and reported that detectors struggle at safe false-positive operating points and generalize poorly to unseen generators and decoding strategies.

Recent work by Jabarian and Imas (2025), in a paper covering multiple genres and text lengths, found the strongest commercial detectors performed far better than the tools tested in 2023, with Pangram achieving near-zero false positive and false negative rates on unmodified, raw AI text, while an open-source baseline flagged 30 to 69% of human writing. Van Vlasselaer et al. (2026) tested four tools on controlled texts and 1,163 real master's theses, finding large performance gaps among tools and sensitivity to humanized and hybrid text. Hadra et al. (2026) tested Turnitin and Originality.ai on 192 texts and found macro-accuracies of 0.61 and 0.69, respectively, with near-zero recall on hybrid human-AI texts and significant performance declines based on the length of the text.

Two important features are worthy to consider in this study. First, reported accuracies for commercial detectors range from about 0.6 to better than 0.99 depending on tool, corpus, and year. This implies that at any given moment the deployed tools differ significantly in behavior. Second, every study scores multiple detectors on a common corpus and evaluates each against ground truth. The between-detector comparison, i.e. whether the tools agree with each other on individual texts, remains unexplored, even though the data these studies collect would support it.

3.3 False positives

Since a false positive can trigger a misconduct proceeding, the literature on false positives has been critical. Liang et al. (2023) demonstrated that seven detectors flagged an average of 61% of TOEFL essays by non-native English speakers as AI-generated while performing almost perfectly on essays by US students. This gap was attributed to the lower lexical (wording) and syntactic (structural) variability of L2 writing. Hadra et al. (2026) found related evidence of bias against texts by EFL students. Vendor-reported false-positive rates span two orders of magnitude, from Pangram's figure near 1 in 10,000, to rates above 1 percent for several mainstream tools, and independent estimates frequently exceed vendor reports (Jabarian & Imas, 2025).

Institutional behaviour has tracked these findings unevenly. OpenAI withdrew its own AI text classifier in July 2023, after citing low accuracy. It identified only 26% of AI-written text while mislabelling 9% of human text. Vanderbilt University disabled Turnitin's AI indicator the same year, estimating that at its submission volume in 2022 the tool's claimed 1% false-positive rate would have wrongly implicated roughly 750 student papers per year. Other institutions followed after that. Others retained the tools, typically with guidance that scores should not be the only evidence. None of these policies considered evidence about inter-tool consistency, or whether an accused student's "98% AI" would be scored the same on a different detector used by another institution.

3.4 Theoretical limits

Several studies in the literature examine robustness. Sadasivan et al. (2023) showed that light paraphrasing sharply degrades watermark- and perplexity-based detection, and reported that, as the distributions of human and model text converge, the best achievable detector approaches chance. Krishna et al. (2023) demonstrated that a purpose-built paraphraser (DIPPER) evades a range of detectors, proposing retrieval over generation histories as a defense that is unavailable to third-party tools. RAID's adversarial suite (Dugan et al., 2024) found simple in-text additions such as homoglyphs and repeated whitespace confused many detectors. The present study intentionally excludes adversarial and humanized text from its main conditions. Our question is solely on whether detectors agree on clean material, whether pure human, pure AI, or a hybrid of both.

3.5 Reliability between detectors

Prior work has investigated validity while leaving reliability between instruments unquantified. The classical test theory treats reliability as a maximum on usable validity, and fields that rely on discrete, categorical judgments require demonstrated inter-rater agreement before deploying raters with consequences. Benchmarks such as Landis and Koch's (1977) bands for kappa exist because "percent agreement" alone is inflated by chance. Weber-Wulff et al. (2023) note cases where tools split on the same text, and practitioner comparisons show the same essay scoring near 0 on one tool and near 100 on another. Elkhatat et al. (2023) document inter-tool inconsistency on control texts. But no study has quantified inter-detector agreement with standard reliability statistics, per authorship condition, at the verdict and score levels, which is the primary aim of this paper.

---

4. Methods

4.1 Corpus

The corpus used in this study contains 90 academic research snippets in the range of 280 to 400 words. Thirty were purely human written, 30 were AI generated, and 30 were a hybrid of both together. For each of the three categories, the 30 texts discussed distinct research topics from four disciplines (8 STEM, 7 medicine and health, 8 social sciences, and 7 humanities).

Human condition. Thirty texts were extracted from the introduction or discussion sections of open-access journal articles published before January 2022, that is, before any public access to ChatGPT-class models, and so their human provenance is verifiable by publication date. All sources carry CC BY licences, and the DOI, journal, publication year, licence, and source section of every passage are recorded in the released corpus manifest. Sources include PLOS

journals, Europe PMC full-text articles, and other CC BY venues, with each passage trimmed to the target length at sentence boundaries.

AI condition. Thirty texts were generated one-shot on matched topics by five current AI model families (six texts per family): DeepSeek (deepseek-v4-pro), Mistral (mistral-large-2512), OpenAI (gpt-5.6-luna), Anthropic (claude-sonnet-5), and Google (gemini-3.1-pro-preview). Each model received the same plain prompt: "*write a passage of about 350 words that could appear in the introduction section of an academic research paper on the given topic, in a formal academic tone with a few APA-style in-text citations, with no title, headings, or reference list*". All generations used default settings with no temperature or sampling overrides, no system prompts beyond the instruction, no humanization, no human involvement, and no adversarial prompting. Outputs were accepted as AI generated in their raw form. Model-invented citations were retained, as is normal for one-shot generation. Families were assigned to topics by a fixed rotation approach.

Hybrid condition. Each of the 30 human-written texts was split at sentence boundaries into an opening (roughly the first 33 percent of words) and a closing (roughly the last 33 percent). An AI model from a different family than the topic's AI-condition text wrote a connecting "middle section" of about 150 words in the same formal writing style, and the stored hybrid text is the concatenation: human opening, AI generated middle section, and human closing. The AI part of each hybrid text is recorded per text and averaged about one third of the words. This partitioning simulates the common real-world case of a human writer modifying an AI generated text, or drafting a human written text and then supplementing it with AI generated text. The full corpus, generation protocol, and per-text provenance are made publicly available with the score matrix.

4.2 Detectors and scoring

Nine commercial detectors were selected based on their market presence in academic and publishing workflows. Turnitin's AI checker, GPTZero, Originality.ai, Pangram, Copyleaks, ZeroGPT, Winston AI, Sapling, and QuillBot's AI detector. All 810 text-detector scans were performed on 2026-08-18, so every tool rated the same corpus within a single collection timeframe at whatever model version the vendor was serving during the time. Each text was scanned once per detector. The design measures inter-tool agreement on single verdicts, which is what an instructor uses as a verdict in practice.

Tools report on different scales. Some return the probability that a text is AI-generated, some the estimated percentage of AI-written text, and some a discrete, categorical label with a supporting score. All outputs were mapped onto a common 0 to 100 "AI-score" scale using each vendor's own primary quantity: the AI percentage where the tool reports one (Turnitin, Pangram, ZeroGPT, Sapling, QuillBot, Copyleaks, Originality.ai); the number of sentences flagged as AI for GPTZero's sentence-level output; and 100 minus the human score for the case of Winston. If a tool only had a categorical verdict of human or AI, the verdict was

recorded at the scale endpoints. The tools' own categorical labels (human, mixed, or AI) were recorded if its interface displayed them.

4.3 Agreement analysis

Since vendors do not disclose their internal decision thresholds, the scores were binarized at a uniform 50% for the primary verdict-level analysis (a text is "flagged" if its AI score >= 50), and the entire analysis was recomputed at every threshold from 5 to 95 in steps of 5 to verify that no conclusion depends on the cutoff. Verdict-level agreement was measured with pairwise Cohen's kappa between all 36 detector pairs, Fleiss' kappa, and Krippendorff's alpha; and score-level agreement with Krippendorff's alpha (interval) and Spearman rank correlations. Percent agreement and prevalence-adjusted bias-adjusted kappa (PABAK; Byrt et al., 1993) were also reported, since chance-corrected coefficients are known to collapse in near-homogeneous strata, which applies to the pure-text conditions. Per-text disagreement was summarized by the score range (maximum minus minimum across the nine tools) and by the contradiction rate, or the share of texts receiving a sub-threshold and a supra-threshold verdict from different tools. Confidence intervals are bootstrap percentile intervals over 2,000 text resamples. Secondary accuracy analyses (false-positive and false-negative rates against ground truth) were reported for comparability with prior benchmarks. All of the statistics were computed by deterministic scripts.

5. Results

5.1 Score distributions

Figure F3 shows each raw score by detector and condition. Table 1 provides an overall summary. For the two pure (Human and AI) conditions, the nine tools behave almost identically. Human passages were near 0 (mean 2.5) and pure AI passages were near 100 (mean 97.1). Sapling produced high scores on several human passages (five at 70 or above, and seven between 12 and 35). ZeroGPT missed three AI texts, with scores of 0, 44, and 47. Winston scored one AI text 0 that seven other tools scored 100.

The hybrid condition shows significantly higher disagreement. The same 30 texts, each containing roughly one-third raw AI text, received per-tool mean scores ranging from 0 (Copyleaks scored every hybrid 0) to 72 (Sapling). Between those two extremes, Pangram (mean 33) and Turnitin (mean 28) tracked the true AI proportion; GPTZero labelled 26 of the 30 hybrids "mixed" and 4 "human"; Originality.ai scored them half human and half AI, scoring 15 texts 0 and the other 15 texts 100; and Winston was almost perfectly bimodal, placing 12 hybrid texts at 73 or over, and 17 at 26 or under.

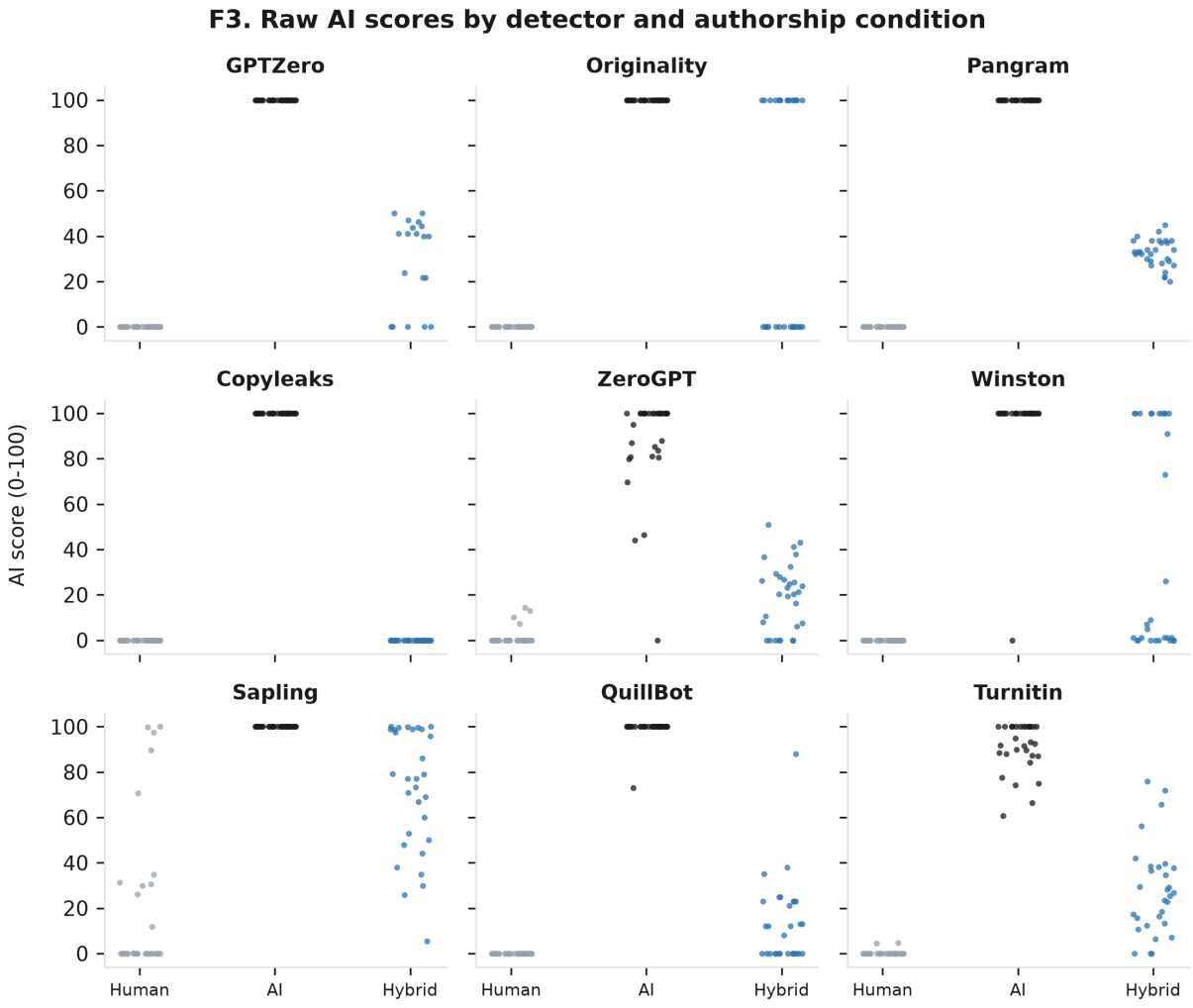

Figure F3. Raw AI scores by detector and authorship condition.

Table 1. Mean / median / min-max AI score by detector and condition.

Detector	Human	AI	Hybrid
GPTZero	0 / 0 / 0-0	100 / 100 / 100-100	29 / 40 / 0-50
Originality.ai	0 / 0 / 0-0	100 / 100 / 100-100	50 / 100 / 0-100
Pangram	0 / 0 / 0-0	100 / 100 / 100-100	33 / 33 / 20-45
Copyleaks	0 / 0 / 0-0	100 / 100 / 100-100	0 / 0 / 0-0
ZeroGPT	1.5 / 0 / 0-14.5	87.4 / 100 / 0-100	19.3 / 21.2 / 0-50.8
Winston	0 / 0 / 0-0	96.7 / 100 / 0-100	41.9 / 7 / 0-100
Sapling	20.7 / 0 / 0-100	100 / 100 / 100-100	71.8 / 77 / 5.4-100
QuillBot	0 / 0 / 0-0	99.1 / 100 / 73-100	13.1 / 12 / 0-88
Turnitin	0.3 / 0 / 0-4.7	91.1 / 93.2 / 60.8-100	28 / 26.8 / 0-76

5.2 Inter-detector agreement

Across the full corpus, agreement statistics were in the range that reliability conventions label substantial: Krippendorff's alpha 0.71 (95% CI 0.61 to 0.78) on binary verdicts and 0.76 (95% CI 0.68 to 0.83) on raw scores, Fleiss' kappa 0.78, mean pairwise Cohen's kappa 0.73 (95% CI 0.66 to 0.80), and mean pairwise percent agreement 86 percent (PABAK 0.73). Score-level Spearman correlations between tools are uniformly high (0.68 to 0.94), and this confirms that all of the nine detector tools separate the easy extremes of the corpus in the same direction.

Within the human condition, mean pairwise percent agreement is 96% (PABAK 0.93), and within the AI condition, 97% (PABAK 0.94). Within the hybrid (human + AI) condition, percent agreement dropped to 64%, PABAK to 0.28, and mean pairwise Cohen's kappa to 0.02, which is chance-level agreement. On texts that mix human and AI writing, knowing one detector's verdict provides essentially no information about what another detector will score.

Figure F1 shows the pairwise structure. Pangram and Copyleaks agreed perfectly at the verdict level (kappa 1.00) since both are extreme-value tools on this corpus, and a cluster of GPTZero, Pangram, Copyleaks, QuillBot, and Turnitin agrees at kappa 0.87 to 0.97. Sapling is the outlier, agreeing with every other tool at kappa 0.36 to 0.62, and Originality.ai and Winston had an intermediate band (0.58 to 0.75) based on their all-or-nothing behavior on hybrids.

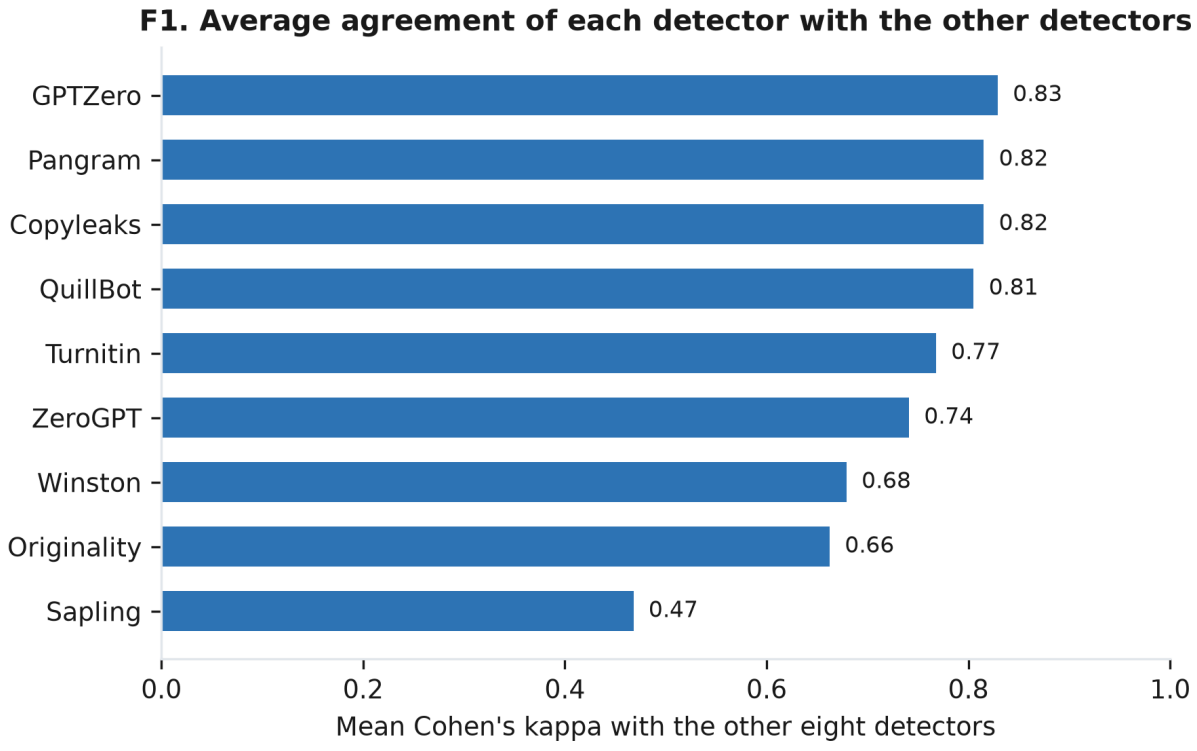

Figure F1. Average agreement of each detector with the other eight detectors (Cohen's kappa).

5.3 Contradictory verdicts and per-text spread

Across the corpus, 37 of 90 texts (41%; 95% CI 31 to 51) received both a “human” verdict and an “AI” verdict from different tools at the 50% threshold, and 30 of 90 (33%) show strong contradictions, with at least one tool at or below 10 and another at or above 90 on the same text. Figure F2 presents a plot of each text’s score range. The mean range is 45 points overall, but 22 points in each pure condition vs. 93 points in the hybrid condition, where the median range is the 100-point scale.

The contradiction rate was on the hybrid condition. In total, 28 of 30 hybrid texts (93%) received contradictory verdicts, against 5 of 30 human and 4 of 30 AI texts. One hybrid passage in the social sciences group (SS1) was scored 0 by Originality.ai and Copyleaks, 7.6 by ZeroGPT, 7 by Winston (verdict: human), 13 by QuillBot, 27 by Turnitin, 34 by Pangram, “mixed” with 41 percent of sentences flagged by GPTZero, and 95.7 by Sapling. A student submitting this text would be cleared or accused, depending entirely on which tool their institution uses for AI detection.

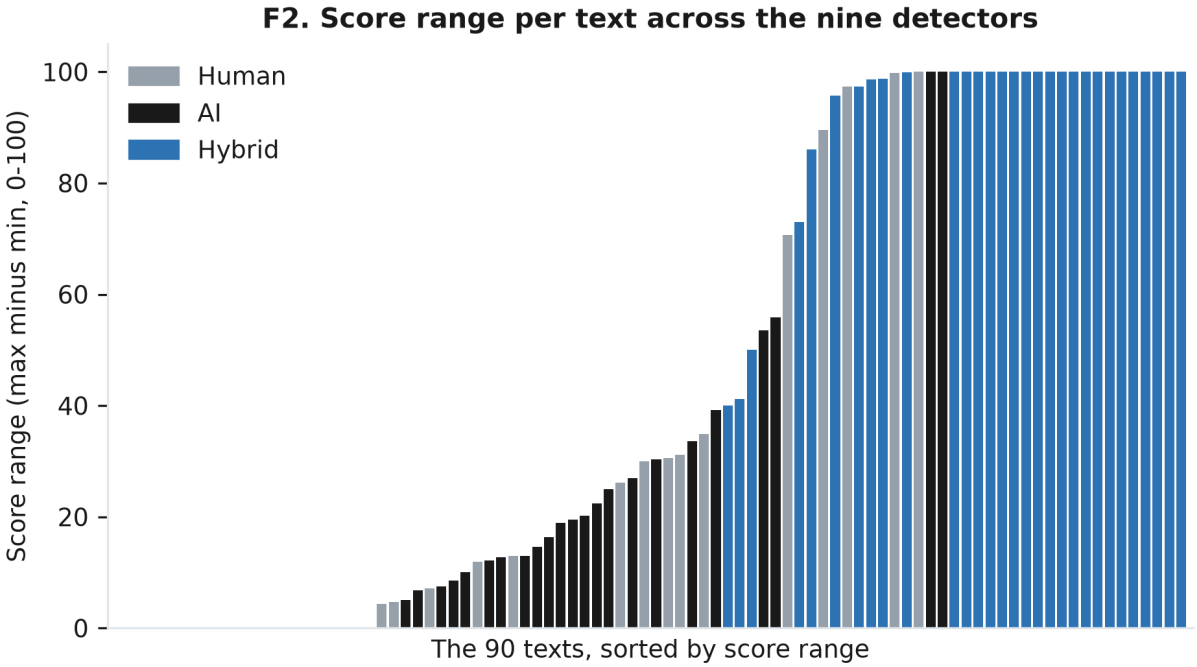

Figure F2. Score range per text (maximum minus minimum across the nine detectors), sorted, colored by condition.

5.4 Accuracy against ground truth

For comparability with prior benchmarks: at the 50% threshold, seven of the nine tools yielded zero false positives on the 30 human passages. Sapling flagged 5 of 30 (16.7%), and no other tool flagged any. False negatives on pure AI text were similarly rare. ZeroGPT missed 3 of 30 and Winston 1 of 30, while the other seven tools missed none. These figures on clean text are consistent with the strongest recent benchmarks (Jabarian & Imas, 2025) and demonstrate that the corpus has no unusual difficulty. The disagreement documented above therefore cannot be attributed to weak tools failing on hard text. The same instruments that

are individually near-perfect on pure (human or AI) texts diverge to chance-level mutual agreement when they are combined, e.g., human-modified or -paraphrased AI text. Treating "contains AI" as the ground truth for hybrids, per-tool hybrid detection ranges from 0 of 30 flagged (Copyleaks) to 23 of 30 (Sapling), with most tools between 1 and 15 (see Figure F5).

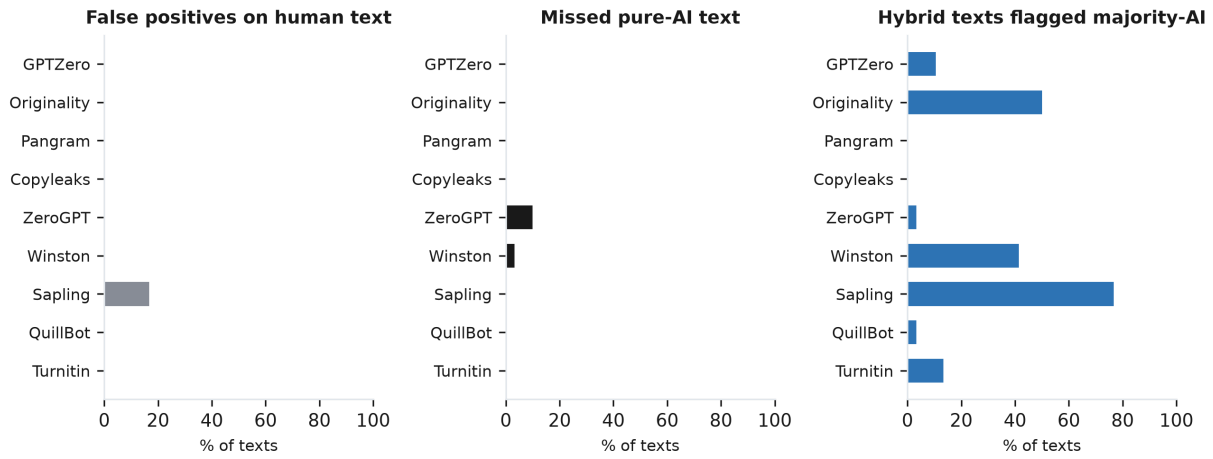

Figure F5. False positives, false negatives, and hybrid flag rates by detector.

5.5 Thresholds and rankings

Recomputing the entire verdict-level analysis at each uniform threshold from 5 to 95 changes Krippendorff's alpha only between 0.61 and 0.73 (Figure F4), with the maximum near the 50 to 80 band used for the primary analysis. No threshold choice fixes agreement, so the disagreement reflects genuinely different text-level judgments, and not the different cutoffs applied to similar scores. In line with this, rank correlations are high globally, but decline within the hybrid condition, where tools order the same 30 texts in materially different ways. Pure-AI unanimity held across all five AI model families and all four discipline groups (the few pure-AI misses were detector-specific, not AI family-specific), and hybrid disagreement appeared across each discipline group, at six texts per family, and seven to eight per discipline, these breakdowns are reported as descriptives only.

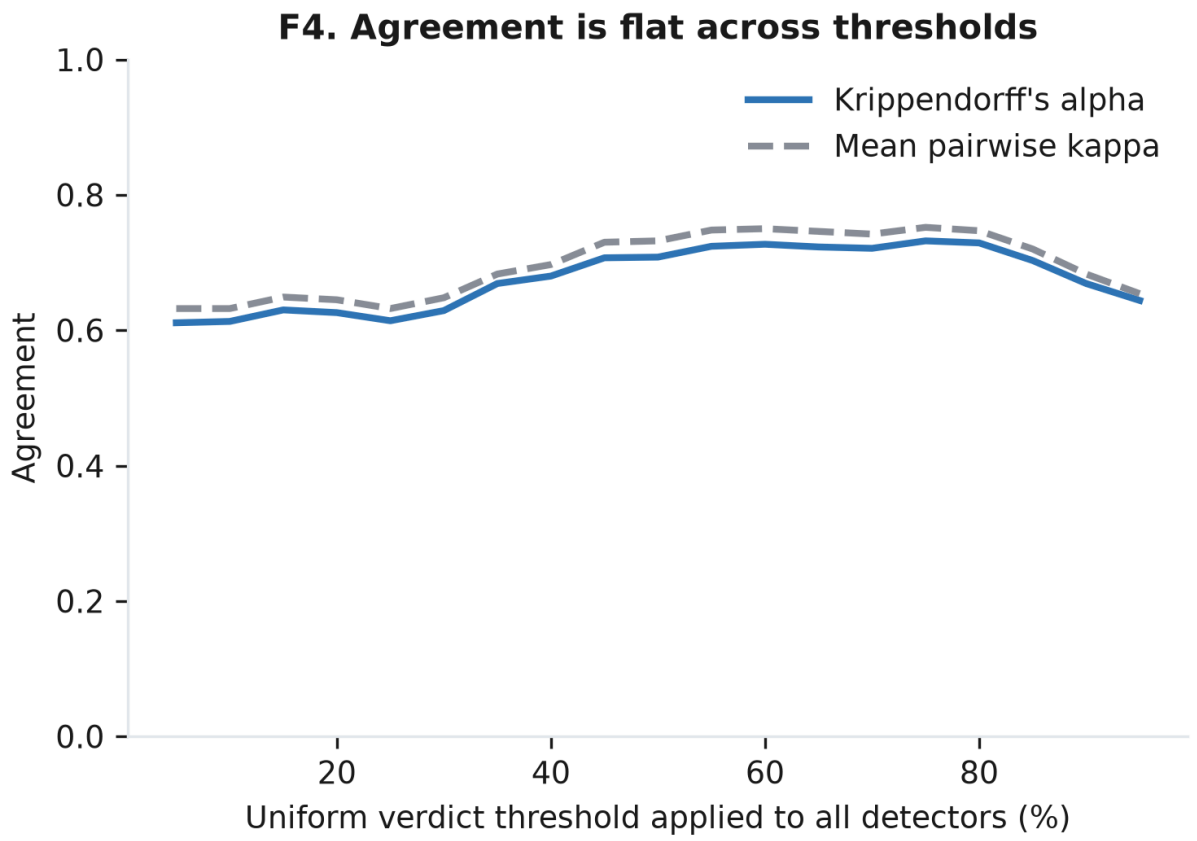

Figure F4. Agreement is flat across thresholds.

6. Discussion

6.1 Two rules, one average

At first glance, this study concludes that commercial AI detection is reliable. An alpha value of 0.71 clears conventional bars for exploratory research. The number is revealed as the average over two regimes that should never be averaged. On raw, unmodified text, nine independently built commercial tools behave close to interchangeably, which is considered progress over the tools tested in 2023 (Weber-Wulff et al., 2023). On text that mixes human and AI authorship, however, the same nine tools agree with each other at chance level (pairwise kappa 0.02). Every practical claim a detector score supports therefore depends on a fact the score does not reveal, and that is whether the text belongs to the easy “pure text” setting or the hard “hybrid” one. Hybrid human-AI text, a human written text that is AI-assisted, is the common way generative tools are actually used in writing (Van Vlasselaer et al., 2026), and it is precisely where the verdict becomes a property of the tool rather than of the text.

6.2 Why the tools diverge

The threshold sweep rules out the most benign explanation. If tools ranked texts alike and drew lines in different places, then some uniform threshold would restore agreement, but none does. The distributions in Figure F3 point to the real cause. On partial authorship, the vendors do not answer the same question. Pangram, Turnitin, and GPTZero's sentence-level output behave like proportion estimators, correctly returning values near the true one-third AI share. Originality.ai and Winston behave like binary classifiers forced to pick a side, and they pick sides roughly at chance across texts. Copyleaks behaves like a conservative whole-text classifier and returns 0 on every hybrid. Sapling behaves like a sensitive fragment detector and saturates. Each behaviour is internally coherent. The vendors' interfaces do not distinguish them, and all render as an "AI score" out of 100. The products measure different constructs and report them on the same-looking metric. This is a construct-validity failure at the category level, which is invisible to any single-tool accuracy study and only measurable when the tools are treated as raters of common material, as done in this study.

6.3 Implications for institutions

Three practical consequences for institutions merit further consideration. First, a single detector score on a text with any mixed authorship should not function as evidence, since the counterfactual score from another tool is close to uninformative about it. At the observed hybrid contradiction rate of 93%, a "98% AI" from one licensed tool coexists with a "human" verdict from another more often than not. Second, second-opinion protocols diverge. If institutions accept a second tool's flag as validation, symmetry requires them to accept a second tool's "human" score as a second verdict, and at chance-level agreement, there is always a chance their scores would contradict each other. Third, the finding gives quantitative support to the institutional withdrawal already occurring, where universities have downgraded to only "advisory" status. Our results suggest a sharper policy line. Detector outputs on mixed (human and AI) authorship texts are tool-specific opinions about an undisclosed construct, and procedures should treat them accordingly. None of this argues that detection is useless. On pure AI text, the tools are in line with one another almost perfectly, and unanimity across several independent detectors is itself informative. On the other hand, the results obtained reject the evidentiary use of any single score on human writing that is AI assisted.

6.4 Limitations

This study comes with several limitations. First, all scans used the tiers of each product available to us during the research. Vendors update models continuously, so the numbers are for that time period only, though the design (agreement, not accuracy) is less exposed to version difference than benchmark studies. Second, n = 30 per condition supports the overall and condition-level statistics, but only descriptive family- and discipline-level breakdowns. Third, the corpus is English-language academic research in the range of 280 to 400 words. Agreement may differ for essays, longer texts, or other non-English languages. Fourth, the hybrid condition implements single splice structure (about a third human part, a third AI

middle part, and a third human part). Other mixing patterns, including AI-assisted editing of human drafts, may behave differently and are deferred for future work. Fifth, the uniform binarization at 50% is a research setting, and not any vendor's official threshold. The threshold sweep shows that no alternative uniform cutoff changes the conclusions. But vendor-specific undisclosed thresholds could change individual pairwise numbers.

7. Conclusion

Nine commercial AI text detectors, scored as independent raters of 90 academic texts, agree almost perfectly that pure human text is human and pure AI text is AI generated. On texts that mix the two (human + AI), their agreement decreases to chance. About 93% of hybrid texts received both a "human" and an "AI" verdict, while the median hybrid text was in the 0-to-100 score range across all AI detectors. The tools report different constructs on the same metric. For the increasing share of real writing that is partially AI-assisted, which detector an institution licenses matters more than what the writer actually wrote, and a single detector score should not be treated as evidence. We make the complete 90-text corpus with per-text provenance, the full score matrix, as well as all the analysis code publicly available for further investigation, with the results obtained in this study as a baseline.

8. Data availability

The corpus (30 human-written texts with DOI-level provenance, 30 AI-generated texts with model and prompt records, 30 hybrid (human + AI) texts with splicing metadata), the detector-by-text score matrix, the generation protocol, and deterministic analysis scripts are archived on Zenodo and available at textpulse.ai/research.

9. References

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46*(5), 423-429.

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic. *Journal of Applied Learning and Teaching, 6*(2).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Coley, M. (2023). Guidance on AI detection and why we're disabling Turnitin's AI detector. Vanderbilt University.

Dugan, L., et al. (2024). RAID: A shared benchmark for robust evaluation of machine-generated text detectors. *Proceedings of ACL 2024*, 12463-12492.

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity, 19*, 17.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382.

Hadra, M., Cambridge, K., & Mesbah, M. (2026). Evaluating the accuracy and reliability of AI content detectors in academic contexts. *International Journal for Educational Integrity*.

Jabarian, B., & Imas, A. (2025). Artificial writing and automated detection. *NBER Working Paper 34223*.

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (4th ed.). Sage.

Krishna, K., et al. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. *NeurIPS 2023*.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 100779.

OpenAI. (2023). New AI classifier for indicating AI-written text. Update of July 2023 announcing the classifier's withdrawal.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? *arXiv:2303.11156*.

Van Vlasselaer, M., Van Droogenbroeck, F., & Spruyt, B. (2026). Who wrote this? Evaluating the reliability of AI detection tools in higher education. *International Journal for Educational Integrity*.

Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing. *Open Information Science, 7*(1).

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*, 26.