Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Vernacular Remote Control: Adversarial Robustness Testing of Six Commercial AI Text Detectors on Provenance-Guaranteed Pre-AI Prose

James Guppenberger

jguppenb@purdue.edu

Purdue University West Lafayette

Research Article

Keywords: AI text detection, academic integrity, false positives, adversarial robustness, vernacular analysis, GPTZero, Pangram, Originality.ai, Sapling

Posted Date: June 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9596240/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Commercial artificial intelligence (AI) text detectors are increasingly used in academic-integrity proceedings, yet their validity rests on the assumption that they measure authorship rather than vernacular style. We tested this assumption directly by scoring 50 provenance-guaranteed pre-AI prose excerpts (1724 to 2002) across six commercial detector configurations, then applying a deterministic rule-based perturbation pipeline implementing 11 adversarial conditions to produce 600 paired observations per detector. The pre-registration was committed publicly before scoring began (Bonferroni alpha = 0.000758 across 66 paired tests). On baseline pre-AI human prose, three detectors produced false-positive rates between 28.0% and 64.6% at the 0.5 threshold (Sapling full 64.6%, Sapling clipped 56.0%, Originality.ai academic 32.0%), while three produced 0.0% false positives (GPTZero, Pangram fraction, Pangram avg). A simple structural perturbation (paragraph reformatting with lists, topic sentences, and discourse markers) shifted GPTZero scores by Cohen's d = 1.25 and pushed 25.0% of confirmed-human spans across the 0.5 threshold, replicating prior findings on Tier A vulnerability. The same perturbation produced inverted responses across vendors: paragraph reformatting raised GPTZero and Pangram avg scores while lowering Originality.ai scores by Cohen's d = -0.76. Cross-vendor correlations were near zero (mean absolute Pearson r = 0.20 across pairwise comparisons), and the two Pangram output configurations (fraction and avg) diverged on 7 of 12 conditions despite originating from the same API call. Pilot replication of the Originality.ai column produced Pearson r = 0.9901 across condition means, demonstrating methodological reproducibility. These findings indicate that commercial AI text detectors do not measure authorship but rather vendor-specific stylistic preferences, that the choice of detector materially determines whether human writing is flagged as AI-generated, and that adversarial third-party perturbation of authentic human prose can deterministically push the most reliable detectors past institutional decision thresholds. We provide tier-based detector-selection criteria for institutions and outline the responsible-disclosure pathway pursued for this study.

1. Introduction

The deployment of generative large language models (LLMs) in higher education has produced rapid institutional adoption of commercial text detectors marketed as classifiers of human versus AI authorship. These tools are used to flag suspected academic-integrity violations, support disciplinary proceedings, and inform admissions screening. The validity of any such use depends on a single assumption: that the detector measures authorship rather than incidental properties of the text. If detectors instead respond to vernacular features (sentence length distributions, transition phrasing, structural conventions), then vernacular variation alone can drive false positives, and identical perturbations of authentic human writing can produce arbitrary AI-attribution outcomes.

Prior work has established that this assumption is fragile. Weber-Wulff et al. [1] tested 14 detection tools and reported that none met the threshold of 80% accuracy required for institutional reliance, with rates of false positives and false negatives sufficient to compromise adjudication. Liang et al. [2] demonstrated that seven widely-used detectors misclassified more than half of TOEFL essays written by non-native

English speakers as AI-generated while accurately classifying US eighth-grade native essays, attributing the bias to detector reliance on perplexity and linguistic constraint as proxies for AI authorship. Perkins et al. [3] tested adversarial attacks against seven detectors and observed that simple obfuscation techniques (typo insertion, homoglyph substitution, paraphrasing) substantially reduced detection accuracy across the panel.

These prior studies establish a pattern: commercial detectors respond to vernacular features that correlate imperfectly with actual authorship, and this correlation can be exploited or accidentally violated by writers whose style does not match the implicit reference distribution. The present study extends this line of inquiry along three axes that prior work has not addressed jointly.

First, we use a corpus that is provenance-guaranteed pre-AI by construction. All 50 baseline spans are drawn from texts published between 1724 and 2002, predating any plausible large-language-model contamination at the level of authorship. This eliminates the major confound in prior detector-validation work, where contemporary human samples may themselves have been AI-assisted at the writing or revision stage. Pre-AI authorship is a logical impossibility under modern LLM technology and provides a clean ground truth.

Second, we test perturbations that are mechanical and reproducible rather than human-rewritten or LLM-generated. The perturbation pipeline implements 11 deterministic transformations grouped into lexical, syntactic, punctuation-level, and structural classes, plus a maximum-trigger combination and an inverse minimum-trigger condition. Every transformation is auditable, and identical inputs produce identical outputs. This lets us separate detector responses to vernacular features from confounds introduced by human or LLM rewriting.

Third, we test six detector configurations across five commercial vendors (GPTZero, Pangram, Originality.ai, Sapling clipped, Sapling full, with Pangram producing two distinct output fields from a single API call), allowing direct vendor comparison and cross-detector divergence analysis. The panel composition was pre-registered, and ZeroGPT was excluded a priori based on documented quality issues that would have introduced confounding noise (see Methods Section 2.7).

The study addresses five hypotheses pre-registered before any scoring took place:

H1 (Tier B replication): trigger-high perturbations will raise mean AI-probability scores on Sapling clipped, Sapling full, Originality.ai academic, and Pangram fraction relative to baseline.

H2 (Tier A breakthrough): the maximum-trigger combinatorial condition will move GPTZero and Pangram avg scores by enough to push at least 25% of spans across the 0.5 threshold from a baseline false-positive rate below 1%.

H3 (Bidirectional control): the trigger-low condition will lower mean AI-probability on at least one detector relative to baseline.

H4 (Cross-detector divergence): a significant detector by condition interaction will be present in a mixed-effects model.

H5 (Pangram fraction vs avg): the two Pangram output configurations will diverge in their response to perturbations, demonstrating that the choice of API output field materially affects institutional false-positive risk.

The threat model motivating the perturbation analysis is third-party adversarial editing, in which a bad-faith actor (a vengeful peer, a competitor, a stalker) modifies an authentic human document to trigger an AI-detection alert against a target. This is not a hypothetical scenario. Recent litigation including *Doe v. Yale University* [4] and similar cases has placed institutions under direct legal exposure when academic-integrity decisions rest on detector outputs that may have been adversarially induced. We do not develop novel attack capability; the perturbation rules implemented here are documented stylistic features already present in the AI-text linguistics literature, and we publish them as a defensive contribution to support institutional detector-selection criteria. Results were shared with detector vendors before publication under a 90-day responsible-disclosure window.

2. Methods

2.1 Corpus construction

The baseline corpus comprises 50 prose excerpts of 250 words each, drawn from 28 source works published between 1724 and 2002 (mean year = 1910). All sources were selected for two properties: provenance-guaranteed pre-AI authorship and stylistic diversity across genres, periods, and registers. The full source list appears in Table 1.

Table 1
Baseline corpus composition (50 spans, 28 source works)

Author	Title	Year	Spans
Daniel Defoe	Roxana	1724	2
Constitutional Convention	U.S. Constitution	1787	1
Hamilton, Madison, Jay	The Federalist Papers	1788	2
Jane Austen	Pride and Prejudice	1813	2
Mary Shelley	Frankenstein	1818	2
Edgar Allan Poe	The Tell-Tale Heart	1843	1
Charles Darwin	On the Origin of Species	1859	2
Charles Dickens	A Tale of Two Cities	1859	2
Mark Twain	Adventures of Huckleberry Finn	1884	2
Arthur Conan Doyle	The Adventures of Sherlock Holmes	1892	2
Bram Stoker	Dracula	1897	2
H.G. Wells	The War of the Worlds	1898	2
Franz Kafka	The Metamorphosis	1915	1
H.P. Lovecraft	Selected works (1917 to 1937)	ca. 1920	1
Virginia Woolf	Mrs Dalloway	1925	1
F. Scott Fitzgerald	The Great Gatsby	1925	1
Isaac Asimov	Liar!	1941	3
Isaac Asimov	I, Robot	1950	3
Isaac Asimov	Youth	1952	2
Martin Caidin	Cyborg novels (1960s to 1980s)	1972	1
Robert M. Pirsig	Zen and the Art of Motorcycle Maintenance	1974	1
Isaac Asimov	A Choice of Catastrophes	1979	2
Douglas Adams	Hitchhiker's Guide / Salmon of Doubt	1979	1
Isaac Asimov	In the Beginning	1981	3
Isaac Asimov	Nemesis	1989	2
Isaac Asimov	The Complete Stories Vol. 1	1990	3
Isaac Asimov	Asimov's Guide to Earth and Space	1991	2

Author	Title	Year	Spans
Christian Cameron	Washington and Caesar	2002	1

Stratification followed a pre-specified table: 20 spans from Asimov works (8 source titles spanning fiction and non-fiction) and 30 spans from non-Asimov sources. Stephen Hawking's works were intentionally excluded from this corpus and reserved for a separate planned study on assistive-technology authorship attribution, given the methodological complexity introduced by Hawking's reliance on predictive-text systems from 1986 forward.

Each 250-word span was extracted at deterministic paragraph offsets (15%, 35%, 60%, 85% of document length, depending on the number of spans drawn from each source). The reproducible random seed used hash(safe_id) so the extraction is exactly replicable from the public-domain source texts. All spans began at paragraph boundaries and excluded section headers, table-of-contents artifacts, footnote markers, and Project Gutenberg [5] metadata. The 50-span corpus has SHA-256 fingerprint and was committed to the pre-registration document before any scoring began.

A pilot study (run identifier vrc_v2_originality_20260501T012611Z) was executed earlier on the same day using a different paragraph-offset pattern (5%, 25%, 50%, 75%) to produce an independent 50-span corpus from the same 28 source works. The pilot scored only Originality.ai and is reported in Section 3.2 as a methodological replication check.

2.2 Perturbation pipeline

Each of the 50 baseline spans was transformed through 11 deterministic perturbation conditions, producing 12 conditions total (including the unperturbed baseline) and 600 spans of analytical data per detector. The pipeline is implemented in Python using NLTK and spaCy (en_core_web_sm) for tokenization, part-of-speech tagging, and dependency parsing. No large language model is involved at any stage of the perturbation.

The 32 individual transformation rules are organized into six classes. Lexical rules (L1 to L7) include Latinate substitution from a curated dictionary of 328 word pairs (replacing Anglo-Saxon roots with Latinate equivalents, e.g., "use" to "utilize", "show" to "demonstrate"), hedge insertion (adding modifiers such as "somewhat," "relatively"), generic transition opening (sentence-initial "Furthermore," "Moreover," "Notably"), contraction expansion ("don't" to "do not"), Zipf-rank-based vocabulary flattening, article over-determination, and adverb proliferation. Syntactic rules (S1 to S6) target sentence-length equalization to a target window of 18 plus or minus 2 words, subject-verb-object linearization, parallel-structure injection, subordination flattening, passive-to-active conversion, and modal-verb regularization. Punctuation rules (P1 to P7) include em-dash elimination, smart-quote normalization, semicolon suppression, Oxford-comma regularization, ellipsis removal, parenthetical flattening, and exclamation-point suppression. Paragraph-level rules (PG1 to PG6) include bulleted-list injection, topic-sentence regularization, conclusion-sentence regularization, paragraph-length equalization, heading injection, and discourse-marker layering ("First," "Second," "Finally").

Combinatorial conditions (C1 to C6) compose multiple rule classes simultaneously. C1 combines Latinate substitution with sentence-length equalization. C2 combines generic transitions with topic-sentence regularization and conclusion-sentence regularization. C3 is a punctuation cluster (em-dash elimination, semicolon suppression, ellipsis removal, exclamation suppression). C4 is a structural cluster (bulleted lists, paragraph-length equalization, heading injection, discourse markers). C5 is the maximum-trigger profile, combining lexical, syntactic, punctuation, and structural transformations in a fixed order (L1 to L3 to S1 to S3 to P1 to P3 to PG2 to PG6). C6 is the inverse minimum-trigger profile.

The trigger-low condition (LOW1 to LOW6) inverts the maximum-trigger logic: it amplifies sentence-length variance, injects contractions, inserts idioms and dialect, adds em-dashes, breaks sentences into fragments, and amplifies concrete sensory detail. This condition tests bidirectional controllability of detector scores.

Every rule is constrained to preserve word count within plus or minus 10% of the original; any rule application that exceeds this band is undone and skipped, with the skip recorded in the audit log. Across all 600 perturbed spans, the highest mean Levenshtein edit distance from baseline was 254.14 characters (C5 maximum-trigger, mean output length 268.32 words from 250-word inputs). Skipped-rule counts are documented in the supplementary materials.

2.3 Detector panel

Six detector configurations were scored across five API endpoints. GPTZero [6] using the documents[0].completely_generated_prob output field. Sapling clipped [7] on the first 480 characters of each span (the documented free-tier input limit), using the score field. Sapling full on the entire 250-word span (approximately 1500 characters, requiring a Pro tier account) on the same endpoint and field. Originality.ai academic [8] using aiModelVersion="academic" and the score.ai field. Pangram [9] (using the pangram-sdk Python client) producing two separate detector configurations from a single API call: pangram_fraction reads the response.fraction_ai_content field, while pangram_avg computes the mean of the per-window ai_assistance_score values returned in response.windows .

Detector selection criteria were pre-registered. We required commercial availability, English-language support, documented use in academic-integrity contexts, and an API endpoint suitable for programmatic 250-word inputs. ZeroGPT was considered for inclusion but excluded a priori based on documented quality concerns in the 2026 detector-review literature (false positives on edited human prose, false positives on non-native English writing) that would have introduced confounding noise without contributing unique signal beyond what the remaining six configurations already span. The API budget saved by ZeroGPT exclusion was reallocated to expanded test-retest stability sampling (20 spans across 3 additional repeats, see Section 2.5).

2.4 Statistical analysis plan

The statistical plan was committed to the pre-registration document before any scoring began. Primary inferential tests are paired t-tests comparing baseline against each non-baseline condition, separately

per detector (6 detectors x 11 conditions = 66 paired tests). Bonferroni correction was applied for the 66-test family, producing a corrected alpha threshold of 0.05 / 66 = 0.000758. Effect sizes are reported as paired Cohen's d with 95% bootstrap confidence intervals computed from 10,000 resamples of paired baseline-condition differences. Threshold-crossing analyses report the proportion of spans crossing 0.5, 0.7, and 0.9 thresholds per condition per detector, with two-proportion z-tests (acknowledged as conservative under paired data; see limitations in Section 4.7).

A linear mixed-effects model (statsmodels MixedLM) tested the cross-detector divergence hypothesis using the formula score ~ detector * condition + (1 | span_uid). The model could not include crossed random intercepts for both span and author due to statsmodels' single-grouping limitation; the span-level intercept yielded a singular fit on the six-detector panel, so a baseline_uid grouping was used as a fallback. R/lme4 with crossed random effects would be appropriate for confirmatory replication but was not available in the present analytic environment.

For the H5 Pangram fraction-vs-avg comparison, paired t-tests directly compare the two Pangram output configurations across all 12 conditions, with Bonferroni correction at alpha = 0.05 / 12 = 0.004167.

2.5 Test-retest reliability

A subsample of 20 spans (7 baseline, 7 C5 maximum-trigger, 6 trigger-low; stratified by author group) was scored 3 additional times on each detector, with a 1-hour inter-repeat gap. Cronbach's alpha was computed across the 4-repeat scores per detector. The 1-hour gap was a deliberate compromise; a 12-hour gap would be the rigorous standard but was not feasible within the single-evening execution window. The actual gap is documented in the events log and is acknowledged as a limitation in Section 4.7.

2.6 Pre-registration and audit chain

The pre-registration document (PRE_REGISTRATION.md, 3,532 bytes) was generated programmatically before the first detector API call, including the corpus SHA-256 fingerprint, the full hypothesis set, the statistical plan, and the detector panel definition. The document's SHA-256 hash is. The hash was committed to a public, hash-verifiable timestamp service on 2026-05-01 before scoring began, time-stamping the hypothesis commitment and corpus state. The full event log (events.jsonl) records every phase transition, every preflight test, every heartbeat, and every credit-exhaustion incident with UTC timestamps.

Score data writes followed an incremental-flush protocol: every API response was written to scores_raw.jsonl within one second of receipt and never buffered in memory. The cleaned scores_summary.csv was rebuilt from scores_raw.jsonl by selecting the first successful score per (span_uid, condition, detector, repeat_index) tuple, where success required a non-null extracted score and a null/empty error field. This protocol allowed transparent recovery from API quota exhaustion (described in Section 4.7) without compromising data integrity.

2.7 Pangram credit exhaustion and recovery

During primary scoring, the Pangram API returned HTTP 401 "Insufficient credits" responses on a hard cliff after approximately 100 successful primary calls, persisting through the remainder of the run. Diagnostic analysis (preserved as _PANGRAM_DIAGNOSIS.md in the supplementary materials) confirmed the failure was a billing-quota issue rather than a rate-limit, authentication, or input-format problem. Pangram credits were topped up in two stages, and a Pangram-only resume pass with a tightened circuit breaker (2 consecutive 401 limit, 480 hard call cap) recovered the remaining tuples. Final per-tuple coverage on the Pangram configurations was 100% (660/660), matching coverage on the four other configurations except for 24 tuples on GPTZero and Sapling full that failed with HTTP 400 / JSON parse errors on a specific cluster of Dickens C1-C5 perturbed spans (see Section 4.7 for discussion of this edge-case finding).

The Pangram recovery path was approved post-hoc but does not affect the confirmatory status of the analysis: the pre-registration document and corpus SHA-256 were committed before scoring began, hypotheses were not modified, and the perturbation pipeline was bit-identical to the version used in primary scoring. The recovery added only additional successful API calls against the pre-committed input corpus.

3. Results

Total scoring volume across the six-detector confirmatory panel was 3,300 pre-registered API calls (3,000 primary plus 300 test-retest), executed between 2026-05-01T11:49Z and 2026-05-02T02:33Z. Per-tuple coverage at the cleaned-data stage was 100% (660/660) on Sapling clipped, Originality.ai, Pangram fraction, and Pangram avg, and 96.4% (636/660) on GPTZero and Sapling full. The 24 missing tuples on the latter two detectors clustered on the same five Dickens C1 through C5 perturbed spans, with HTTP 400 and JSON parse errors indicating a parser-level failure on a specific input pattern; this is a finding in itself and is discussed in Section 4.7. All paired tests reported below are based on baseline-condition pairs that have valid scores in both cells, with n indicated per cell.

3.1 Baseline false-positive landscape

The clearest result of the study is the baseline false-positive distribution across the six detector configurations on confirmed pre-AI human prose. Three detectors produced false-positive rates statistically indistinguishable from zero across all three institutional decision thresholds (0.5, 0.7, 0.9). Three detectors produced false-positive rates between 28.0% and 64.6% at the 0.5 threshold. The distribution does not span a continuum; it splits cleanly into two performance tiers, summarized in Table 2 and visualized in Fig. 1.

Table 2
Baseline false-positive rates on confirmed pre-AI prose (n = 50 spans, 1724 to 2002)

Detector	n	Mean score	FPR @ 0.5	FPR @ 0.7	FPR @ 0.9
GPTZero	48	0.0014	0.0%	0.0%	0.0%
Pangram fraction	50	0.0000	0.0%	0.0%	0.0%
Pangram avg	50	0.0181	0.0%	0.0%	0.0%
Originality.ai academic	50	0.3490	32.0%	32.0%	28.0%
Sapling clipped (480c)	50	0.5664	56.0%	54.0%	50.0%
Sapling full (250w)	48	0.6468	64.6%	64.6%	62.5%

This separation is the most consequential institutional finding of the study and is independent of any perturbation analysis. Sapling full classified 64.6% of pre-AI human prose as AI-generated at the 0.5 threshold and 62.5% at the 0.9 threshold. The corresponding rates for Sapling clipped were 56.0% and 50.0%, and for Originality.ai academic 32.0% and 28.0%. By construction, every span in the corpus is human-authored. The Constitution of the United States (1787), Pride and Prejudice (1813), Frankenstein (1818), and the Federalist Papers (1788) are texts whose pre-AI provenance is not in question. A detector that classifies 50% to 65% of such texts as AI-generated is not measuring authorship.

Tier A detectors (GPTZero, Pangram fraction, Pangram avg) produced a sharply different baseline distribution. GPTZero returned 0 of 48 spans above the 0.5 threshold, with mean baseline score 0.0014. Pangram fraction returned exactly zero across all 50 baseline spans, with mean and standard deviation both equal to 0.0000. Pangram avg returned mean 0.0181 with no spans crossing 0.5. These three configurations behave on pre-AI prose as a valid AI-detection system should: a near-zero false-positive rate is a necessary condition for the test to be informative when applied to disputed contemporary submissions.

3.2 Pilot to confirmatory replication

A pilot run of the Originality.ai column on an independent 50-span corpus drawn from the same 28 source works (paragraph offsets 5%, 25%, 50%, 75% rather than the confirmatory 15%, 35%, 60%, 85%) was scored prior to the confirmatory study. The pilot used identical detector configuration and identical perturbation pipeline. Comparison of pilot to confirmatory means across all 12 conditions, shown in Table 3 and Fig. 2, demonstrates strong methodological replication.

Table 3
Pilot to confirmatory replication on Originality.ai academic (12 conditions)

Condition	Pilot n	Pilot mean	Pilot SD	Confirm n	Confirm mean	Confirm SD
baseline	50	0.306	0.413	50	0.349	0.439
lexical_only	50	0.418	0.434	50	0.462	0.432
syntactic_only	50	0.189	0.335	50	0.205	0.347
punctuation_only	50	0.072	0.229	50	0.101	0.274
paragraph_only	50	0.007	0.027	50	0.022	0.090
C1	50	0.244	0.360	50	0.239	0.376
C2	50	0.123	0.297	50	0.156	0.324
C3	50	0.210	0.369	50	0.264	0.410
C4	50	0.325	0.399	50	0.367	0.401
C5	50	0.348	0.424	50	0.422	0.434
C6	50	0.201	0.336	50	0.271	0.405
trigger_low	50	0.001	0.001	50	0.002	0.010

Pearson correlation across the 12 condition means was r = 0.9901 (95% bootstrap CI [0.967, 0.998] from 10,000 resamples). The pilot baseline FPR at threshold 0.5 was 30.0% versus 32.0% in the confirmatory run. The strongest pilot effect (Originality on trigger_low, mean 0.0006) was reproduced in confirmatory (mean 0.0021); both are essentially at floor, and the close agreement across the 12 conditions is unambiguous evidence that the perturbation pipeline produces stable detector responses on different input draws from the same source corpus. We treat this as a successful pre-registered replication and report perturbation analyses on the full confirmatory data only, with pilot data preserved as Supplementary Material 1.

3.3 Perturbation effects: paired t-tests

Across 66 paired t-tests (6 detectors x 11 non-baseline conditions), 9 effects passed the Bonferroni-corrected threshold of alpha = 0.000758. These 9 effects are listed in Table 4, ordered by absolute Cohen's d. Twenty additional effects passed an uncorrected alpha = 0.05 threshold but did not survive Bonferroni correction; these are reported as descriptive in the supplementary materials. Figure 3 visualizes the full per-condition effect landscape across the six-detector panel.

Table 4
Bonferroni-significant paired t-tests (baseline vs. condition, alpha = 0.000758)

Detector	Condition	n pairs	t	p	Mean diff	Cohen's d
Pangram avg	paragraph_only	50	11.17	< 0.001	+ 0.546	+ 1.580
GPTZero	trigger_low	48	10.06	< 0.001	+ 0.201	+ 1.452
GPTZero	paragraph_only	48	8.68	< 0.001	+ 0.382	+ 1.254
Pangram fraction	paragraph_only	50	7.04	< 0.001	+ 0.502	+ 0.996
Originality.ai	trigger_low	50	-5.58	< 0.001	-0.347	-0.788
Originality.ai	paragraph_only	50	-5.38	< 0.001	-0.327	-0.761
Originality.ai	punctuation_only	50	-4.61	< 0.001	-0.248	-0.652
Pangram avg	C2	50	4.44	< 0.001	+ 0.113	+ 0.627
GPTZero	C2	48	3.98	< 0.001	+ 0.131	+ 0.575

The strongest effect in the dataset is Pangram avg under paragraph_only perturbation (paired Cohen's d = + 1.580, mean shift + 0.546). This is followed by GPTZero under the trigger_low condition (d = + 1.452, mean shift + 0.201), GPTZero under paragraph_only (d = + 1.254, mean shift + 0.382), and Pangram fraction under paragraph_only (d = + 0.996, mean shift + 0.502). All four of these large-effect findings involve Tier A detectors and the simplest of the perturbation classes. The maximum-trigger combinatorial profile (C5), which was pre-registered as the strongest manipulation, did not produce significant Tier A movement at the Bonferroni threshold.

The most informative pattern in Table 4 is the directional split. GPTZero and the two Pangram configurations move upward under paragraph_only and trigger_low. Originality.ai academic moves downward under the same paragraph_only and trigger_low conditions, plus punctuation_only. The same perturbation that raises Tier A scores by Cohen's d above 1.0 lowers Originality.ai scores by Cohen's d above 0.5 in the opposite direction. This directional inversion is detailed in Section 3.5.

3.4 Threshold-crossing analysis

Threshold crossings at the 0.5 institutional decision threshold are the most policy-relevant statistic in the dataset. Table 5 reports the proportion of spans crossing 0.5 per detector under selected conditions, with baseline crossings included for comparison. Figure 4 visualizes the full threshold-crossing landscape as a detector-by-condition heatmap; the two-tier separation is unambiguous.

Table 5
Proportion of spans crossing the 0.5 threshold under selected perturbation conditions

Condition	GPTZero	Pangram frac.	Pangram avg	Originality	Sapling clp.	Sapling full
baseline	0.0%	0.0%	0.0%	32.0%	56.0%	64.6%
paragraph_only	25.0%	50.0%	52.0%	2.0%	74.0%	79.2%
C2	6.2%	4.0%	8.0%	14.0%	74.0%	62.5%
punctuation_only	0.0%	0.0%	0.0%	10.0%	38.0%	45.8%
trigger_low	0.0%	0.0%	0.0%	0.0%	70.0%	75.0%
C5 (max trigger)	0.0%	0.0%	0.0%	44.0%	60.0%	50.0%

The pre-registered Tier A breakthrough hypothesis (H2) predicted that the maximum-trigger combinatorial profile (C5) would push at least 25% of spans across the 0.5 threshold on at least one Tier A detector. This specific prediction was not supported. C5 produced 0% threshold crossings on all three Tier A detectors. The hypothesis fails on its narrow pre-registered formulation.

However, the broader question that motivated H2, namely whether vernacular perturbations can move Tier A detectors past institutional decision thresholds at all, is supported on a different perturbation class. The paragraph_only condition (lists, topic sentences, conclusion sentences, paragraph-length normalization, headings, discourse markers) pushed 25.0% of GPTZero spans, 50.0% of Pangram fraction spans, and 52.0% of Pangram avg spans across the 0.5 threshold. The same condition pushed only 2.0% of Originality.ai spans across 0.5 (a downward effect). On all three Tier A detectors, paragraph reformatting alone produces threshold-crossing rates that meet or exceed the H2 25% benchmark. The pre-registered hypothesis was directionally correct but mistakenly identified C5 rather than paragraph_only as the strongest manipulation; we treat H2 as supported on paragraph_only and not supported on C5, with full transparency about the pre-registration mismatch.

Tier B threshold crossings under perturbation are summarized in Table 5 and represent a different phenomenon. Sapling clipped and Sapling full operate from a high baseline (56.0% and 64.6% respectively at 0.5) and most perturbations produce only modest changes around that baseline. Originality.ai operates from a moderate baseline (32.0%) and shows the most condition-sensitive response, dropping to 2.0% under paragraph_only and to 0.0% under trigger_low while rising to 46.0% under lexical_only. This sensitivity to specific perturbation classes is consistent with a detector responding to vernacular features rather than authorship. Figure 5 visualizes the score deltas from baseline for the C1 through C6 combinatorial conditions across all six detectors, and Fig. 6 plots the corresponding perturbation dose-response.

3.5 Cross-detector divergence

The cross-detector divergence hypothesis (H4) predicted a significant detector-by-condition interaction in a mixed-effects model. The fitted model produced 21 of 50 detector-by-condition interaction terms with $p < 0.05$, including six interactions significant at $p < 0.001$ (originality_academic by paragraph_only, $t = -7.46$; originality_academic by trigger_low, $t = -5.46$; originality_academic by lexical_only, $t = +2.77$; sapling_full by paragraph_only, $t = -2.38$; sapling_full by punctuation_only, $t = -2.22$; pangram_avg by trigger_low, $t = -2.53$). The interaction is strongly supported. We note one technical caveat: the model could not include crossed random effects for both span and author due to statsmodels' single-grouping limitation, and the span-level grouping yielded a singular fit on the six-detector panel; the reported model uses baseline_uid as the random-effects grouping. R/lme4 with crossed random effects is recommended for confirmatory replication of the variance decomposition; the fixed-effects pattern is robust to this choice.

The descriptive evidence for cross-detector divergence is more striking than any p-value. The pairwise correlations of detector scores at baseline are reported in Table 6.

Table 6
Cross-detector correlation matrix at baseline (Pearson r, n = 48 to 50)

GPTZero	GPTZero	Pangram frac.	Pangram avg	Originality	Sapling clp.	Sapling full
	1.000	**—**	**0.505**	**0.291**	**0.172**	**0.140**
Pangram frac.	—	1.000	—	—	—	—
Pangram avg	0.505	—	1.000	0.259	0.166	0.162
Originality	0.291	—	0.259	1.000	0.297	-0.053
Sapling clp.	0.172	—	0.166	0.297	1.000	0.634
Sapling full	0.140	—	0.162	-0.053	0.634	1.000

Note. Pangram fraction at baseline has zero variance (all 50 baseline spans returned 0.0000), so correlations involving Pangram fraction are undefined.

Pangram avg correlates with GPTZero at $r = 0.505$, the strongest within-tier correlation in the matrix. Sapling clipped correlates with Sapling full at $r = 0.634$, the strongest within-vendor correlation. Across-vendor correlations are weak: Sapling full correlates with Originality at $r = -0.053$ and with GPTZero at $r = 0.140$. The mean absolute Pearson r across all defined pairwise off-diagonal cells is 0.20. If the six detectors were measuring a common latent property of authorship, we would expect cross-detector correlations clustered above 0.7. The observed pattern is consistent with detectors measuring different vendor-specific stylistic properties that correlate weakly with each other and only modestly with authorship.

The directional inversion observed under paragraph_only and trigger_low is the most consequential expression of this divergence. Under paragraph_only perturbation: GPTZero scores rise by 0.382 (paired d = + 1.254), Pangram avg rises by 0.546 (d = + 1.580), Pangram fraction rises by 0.502 (d = + 0.996), and Originality.ai falls by 0.327 (d = -0.761). Under trigger_low perturbation: GPTZero rises by 0.201 (d = + 1.452) and Originality.ai falls by 0.347 (d = -0.788). The same applied transformation produces statistically large opposite-direction effects on different detector vendors. There is no coherent authorship signal to which all detectors are responding; there are multiple stylistic signals to which different detectors respond differently.

3.6 Pangram fraction versus Pangram avg (H5)

The two Pangram detector configurations originate from a single API call but read different output fields. pangram_fraction reads response.fraction_ai_content (a document-level proportion estimate). pangram_avg reads the mean of the per-window ai_assistance_score values returned in response.windows. Per-condition paired t-tests directly comparing the two configurations across all 12 conditions are reported in Table 7 with Bonferroni correction at alpha = 0.05 / 12 = 0.004167. Figure 8 visualizes the baseline divergence between the two output fields.

Table 7
Pangram fraction vs. Pangram avg paired comparisons (alpha = 0.004167)

Condition	n	Mean diff	t	p	Cohen's d	Sig. Bonf.
baseline	50	+ 0.018	3.31	0.0018	+ 0.468	Yes
lexical_only	50	+ 0.018	3.39	0.0014	+ 0.479	Yes
syntactic_only	50	+ 0.006	9.28	< 0.001	+ 1.312	Yes
punctuation_only	50	+ 0.016	3.36	0.0015	+ 0.475	Yes
paragraph_only	50	+ 0.062	2.35	0.0228	+ 0.332	No
C1	50	+ 0.010	2.85	0.0065	+ 0.402	No
C2	50	+ 0.090	4.12	< 0.001	+ 0.583	Yes
C3	50	+ 0.015	3.30	0.0018	+ 0.466	Yes
C4	50	+ 0.029	4.35	< 0.001	+ 0.616	Yes
C5	50	+ 0.010	4.00	< 0.001	+ 0.565	Yes
C6	50	+ 0.031	4.00	< 0.001	+ 0.566	Yes
trigger_low	50	+ 0.008	4.77	< 0.001	+ 0.674	Yes

Nine of the 12 conditions produce Bonferroni-significant divergence between the two Pangram outputs. The strongest divergence is on the syntactic_only condition (Cohen's d = + 1.312). The differences are small in absolute terms (mean differences range from 0.006 to 0.090), but they are systematic and they

produce different threshold-crossing decisions on selected conditions. Under paragraph_only, Pangram fraction returns mean 0.502 and 50.0% threshold-0.5 crossings, while Pangram avg returns mean 0.564 and 52.0% threshold-0.5 crossings; the institutional decisions are nearly identical here. Under syntactic_only, Pangram fraction returns mean 0.000 and 0.0% crossings, while Pangram avg returns mean 0.006 and 0.0% crossings; again practically identical. But under C2 (transition plus topic-sentence stacking), Pangram fraction returns mean 0.042 and 4.0% crossings, while Pangram avg returns mean 0.132 and 8.0% crossings: different effective false-positive rates from the same API call, depending solely on which output field the institution reads.

H5 is supported. An institution adopting Pangram for academic-integrity screening must specify which output field is consulted, and the choice can change which spans are flagged. Documentation guiding this choice was not visible to us in publicly accessible Pangram materials at the time of writing; this represents an institutional governance gap.

3.7 Test-retest reliability

Test-retest reliability was assessed on a 20-span subsample (7 baseline, 7 C5 maximum-trigger, 6 trigger-low) with three additional repeats per detector. Cronbach's alpha across the four-repeat scores per detector was 1.00 for GPTZero, Sapling clipped, Sapling full, Originality.ai, and Pangram avg. Pangram fraction returned undefined alpha because all baseline-condition responses in the subsample were exactly 0.000 (zero variance); we treat this as effectively perfect reliability with the caveat that the alpha statistic is not informative on a constant series. The high reliability is reassuring given the 1-hour inter-repeat gap (rather than the more rigorous 12-hour gap). Within the limits of the present design, detector outputs are stable on identical inputs. Figure 7 visualizes the per-detector repeat consistency.

3.8 Hypothesis verdicts

Pre-registered hypothesis verdicts at the Bonferroni-corrected confirmatory level are as follows. H1 (Tier B replication on C5): NOT SUPPORTED. The pre-registered C5 manipulation did not produce Bonferroni-significant Tier B effects, although Originality.ai showed strong condition-specific responses on paragraph_only, punctuation_only, and trigger_low (all Bonferroni-significant). H2 (Tier A breakthrough on C5): NOT SUPPORTED on the narrow C5 prediction; SUPPORTED on the broader question of Tier A controllability via paragraph_only on three Tier A detectors. H3 (Bidirectional control via trigger_low): SUPPORTED on Originality.ai (downward) and GPTZero (upward, in the inverse direction to pre-registration intent). H4 (Cross-detector divergence): SUPPORTED via mixed-effects interaction and descriptive correlation evidence. H5 (Pangram fraction vs avg): SUPPORTED on 9 of 12 conditions.

The pattern of supported and not-supported verdicts indicates that the directional structure of the pre-registered hypotheses was largely correct but that the specific manipulation predicted to produce the strongest effects (C5 maximum-trigger combinatorial) was incorrect. The strongest manipulation was paragraph_only, a single rule class. We discuss the implications of this finding for the perturbation taxonomy in Section 4.2.

4. Discussion

This study tested whether commercial AI text detectors measure authorship or vernacular style by scoring 50 provenance-guaranteed pre-AI prose excerpts (1724 to 2002) across six detector configurations, then applying a deterministic rule-based perturbation pipeline to produce 600 paired observations per detector. The findings depart in important ways from the pre-registered predictions, and the departures themselves are informative.

4.1 The two-tier baseline landscape

The most consequential finding of the study is independent of any perturbation analysis. On confirmed pre-AI human prose drawn from sources spanning the U.S. Constitution to the Asimov corpus to Cameron's 2002 historical fiction, three commercial detectors (GPTZero, Pangram fraction, Pangram avg) produced false-positive rates indistinguishable from zero across all institutional decision thresholds, while three (Sapling clipped, Sapling full, Originality.ai academic) produced false-positive rates between 28.0% and 64.6% at the 0.5 threshold. Sapling full classified 62.5% of pre-AI human prose as AI-generated at the 0.9 threshold.

This is not a marginal failure of calibration. Texts whose pre-AI provenance is not in dispute, including the U.S. Constitution, Pride and Prejudice, Frankenstein, A Tale of Two Cities, and the Federalist Papers, are systematically classified as AI-generated by detectors currently deployed in academic-integrity proceedings. The detector cannot be measuring authorship under any reasonable definition of that term. It is measuring some property of the text that correlates with neither authorship nor publication date and that flags large fractions of canonical pre-AI human writing as machine-generated.

The two-tier separation is sharp rather than continuous. There is no detector with intermediate baseline FPR. Either a detector is at floor on pre-AI prose, or it is at 28% to 65%. Institutions selecting a detector for academic-integrity adjudication should treat this binary distinction as the primary selection criterion and treat all other comparative properties (price, throughput, API ergonomics) as secondary.

This finding is consistent with prior detector-validation work but extends it materially. Weber-Wulff et al. [1] reported that no tested detector met an 80% accuracy threshold on a mixed corpus that included some contemporary human writing. Liang et al. [2] reported that seven detectors misclassified more than half of TOEFL essays as AI-generated. The present study removes the contemporary-human confound entirely by using exclusively pre-AI text, eliminating the possibility that the human samples were themselves AI-assisted at the writing or revision stage. Under the strictest possible test of detector validity (text whose AI provenance is logically impossible), the same Tier B failure pattern reappears, and the magnitude is severe.

4.2 Why paragraph_only outperformed C5 on the Tier A breakthrough

The H2 hypothesis pre-registered the maximum-trigger combinatorial profile (C5) as the strongest manipulation, predicting that C5 would push at least 25% of spans across the 0.5 threshold on Tier A detectors from a baseline FPR below 1%. The C5 prediction failed: 0% of GPTZero, Pangram fraction, and Pangram avg spans crossed 0.5 under C5. The paragraph_only condition, applying a single rule class rather than the combinatorial stack, pushed 25.0%, 50.0%, and 52.0% of spans across the 0.5 threshold on the three Tier A detectors respectively. The pre-registered hypothesis was directionally correct (Tier A detectors are vulnerable to vernacular perturbation) but mechanistically wrong (the C5 stack is not the strongest manipulation).

One plausible explanation is perturbation interference. The C5 profile applies eight rule sets in sequence (Latinate substitution, generic transition opening, sentence-length equalization, parallel-structure injection, em-dash elimination, semicolon suppression, topic-sentence regularization, discourse-marker layering). Each rule was selected because it produces an AI-typical stylistic feature, but the combined output may push the text into a register that the detectors recognize as overworked rather than as machine-generated. The paragraph_only condition applies six rules from a single class (bulleted-list injection, topic-sentence regularization, conclusion-sentence regularization, paragraph-length equalization, heading injection, discourse-marker layering) and produces a coherent structural reformat that closely matches the formatting conventions of LLM-generated technical prose without other vernacular signals diluting the structural signature.

If this interpretation is correct, the paper's threat model changes in a useful direction. The simplest single-class perturbation produces the largest effect on the highest-quality detectors, and the perturbation is one that any user with basic word-processor competence can produce in minutes by reformatting a passage with bullets, topic sentences, and discourse markers. This is the threat model that should concern academic-integrity administrators most directly.

4.3 Cross-detector divergence and inverse responses

The cross-detector divergence finding has implications beyond the specific perturbation results. Mean absolute Pearson r across cross-vendor pairwise correlations was 0.20 at baseline. Sapling clipped versus Sapling full correlated at r = 0.634, the strongest within-vendor correlation. Sapling full versus Originality.ai correlated at r = -0.053, essentially uncorrelated and slightly negatively associated. If commercial detectors were measuring the same latent authorship signal, we would expect cross-detector correlations clustered above 0.7. The observed pattern is consistent with each detector measuring a distinct vendor-specific stylistic signature that correlates weakly with authorship.

The directional inversion under paragraph_only and trigger_low formalizes this. Identical applied perturbations move GPTZero and Pangram avg upward by Cohen's d above 1.0 while moving Originality.ai downward by Cohen's d above 0.5. This is not a continuous-magnitude difference, it is a sign reversal. The same intervention makes one detector say more AI and another say less AI, by large effect sizes, on the same pre-AI human text. Detectors are not noisy estimators of a common authorship

signal; they are responding to different stylistic features that are independently correlated with whatever LLM training distribution each vendor used as a reference.

This finding has direct implications for academic-integrity policy. Institutions sometimes use multi-detector concordance as evidence of AI authorship, on the reasoning that two independent detectors agreeing on AI status provides stronger evidence than one detector alone. The cross-detector correlation matrix and the directional inversion finding indicate that two detector outputs can disagree by sign on identical text. Multi-detector concordance is not a free lunch and may produce systematically biased conclusions when one of the detectors has a divergent stylistic profile relative to the others on the document being assessed.

4.4 Pangram fraction versus Pangram avg as institutional governance

The H5 finding (9 of 12 conditions show Bonferroni-significant divergence between Pangram's two output configurations from the same API call) deserves special attention as an institutional governance issue. An institution adopting Pangram for academic-integrity screening must specify which output field is consulted, and the choice changes which spans are flagged. Documentation guiding this choice was not visible in publicly accessible Pangram materials at the time of writing. Two institutions using Pangram nominally for the same purpose may reach different conclusions on identical student work depending solely on which output field their integration reads. This is a direct institutional risk that responsible AI-detection deployment requires policies explicitly addressing.

4.5 Comparison to prior literature

The findings of this study are broadly consistent with the prior literature on detector unreliability while extending it on three dimensions. Weber-Wulff et al. [1] demonstrated that 14 detection tools failed to meet a 80% accuracy threshold on a mixed corpus, with their highest-rated tool reaching only modest accuracy. Our finding that three Tier B detectors produce 28% to 65% baseline FPR on confirmed pre-AI human prose is consistent in direction and stronger in magnitude, likely because our corpus eliminates the contemporary-human confound that may have buoyed prior accuracy estimates.

Liang et al. [2] demonstrated that seven detectors misclassified non-native English writing at rates above 50% and attributed the bias to detector reliance on perplexity and linguistic constraint as proxies for AI authorship. Our finding that paragraph_only perturbation alone (a structural rather than lexical manipulation) pushes 25% to 52% of pre-AI human spans across the 0.5 threshold on Tier A detectors extends the Liang et al. mechanism: detectors are sensitive not only to lexical constraint but also to structural conventions, and a writer who organizes prose with topic sentences and bullet lists may be flagged regardless of native-language status.

Perkins, Roe, et al. [3] tested adversarial obfuscation techniques against seven detectors and reported that simple modifications substantially reduced detection accuracy. Our findings on the trigger_low condition (Cohen's d = + 1.45 on GPTZero, mean shift + 0.20) are notable in this context because the

trigger_low manipulation was designed to make text look more human (burstiness amplification, contractions, fragments, em-dashes, idioms) yet raised AI-probability scores significantly on GPTZero. This is the cross-vendor inversion phenomenon manifesting on a perturbation specifically designed for the opposite effect, indicating that human-style features themselves can trigger AI detection on certain detector vendors.

Sadasivan et al. [10] provided theoretical grounds for skepticism about reliable AI-text detection, demonstrating that recursive paraphrasing attacks break a wide range of detection schemes including watermarking and neural classifiers. The present study contributes empirical evidence in the inverse direction: not that AI text can be made to evade detection, but that confirmed pre-AI human text can be deterministically pushed across detector decision thresholds via mechanical vernacular perturbation. The two findings together describe a detector category that is bidirectionally manipulable on both human and AI inputs.

4.6 Threat model and the Doe v. Yale context

The third-party adversarial threat model motivating this study is not hypothetical. *Doe v. Yale University* [4] arose from the suspension of an Executive MBA student after his final exam was flagged by GPTZero. The plaintiff Thierry Rignol, a French national and non-native English speaker, alleged that Yale's own institutional policies prohibit the use of GPTZero due to documented false-positive rates against non-native English writing, and submitted to the Honor Committee scans showing GPTZero returning '100% probability AI' on academic papers authored by Yale President Peter Salovey from more than 30 years prior. The District Court denied a preliminary injunction (May 5, 2025) and denied pseudonymous-plaintiff status (March 31, 2025), but the underlying breach-of-contract and Title VI discrimination claims remain pending [11].

The case illustrates the institutional exposure created when academic-integrity decisions rest on detector outputs whose validity has not been independently demonstrated. The present study's findings on paragraph_only perturbation indicate that even a detector with a 0% baseline FPR (such as GPTZero) can be deterministically pushed to 25% threshold-crossing rate by simple structural reformatting. A defendant in a future Doe v. Yale-style case could argue that an authentic essay was rendered AI-positive by routine word-processor formatting decisions, that a vengeful peer modified an authentic essay before submission to trigger a false positive, or that the detector's response to vernacular features represents a discriminatory proxy for non-native-speaker status. None of these arguments require expert testimony beyond what this study and its precursors document; the empirical record sufficient to mount such defenses is now publicly available.

The responsible-disclosure pathway adopted for this study reflects this institutional context. The pre-registration document and corpus SHA-256 were committed publicly before scoring, the perturbation pipeline contains no novel attack capability not already documented in the AI-text linguistics literature, and findings were shared with detector vendors before publication under a 90-day disclosure window. We urge institutional adopters to evaluate detector outputs as one input among many, to require

corroborating evidence beyond detector outputs in academic-integrity proceedings, and to adopt explicit policies addressing detector limitations including the directional inversion and Pangram output-field-selection issues documented here.

4.7 Limitations

Five limitations bound the conclusions of this study and warrant explicit acknowledgement.

First, the corpus is English-only and Western-canonical. The 28 source works span 1724 to 2002 but draw primarily from British and American literary traditions plus selected non-fiction (the U.S. Constitution, the Federalist Papers, Darwin's Origin of Species, Asimov's science writing). Findings on detector behavior may not generalize to texts in other languages, to translated works, to non-Western literary traditions, or to texts produced under writing conventions substantially different from those represented here.

Second, the perturbation pipeline operates at the level of paragraph and below; it does not test perturbations of document structure (multi-paragraph reorganization, citation insertion, abstract addition) that would correspond to certain real-world editing scenarios. The perturbation taxonomy is defensible as a starting point but is not exhaustive. Future work should extend the rule set to document-level transformations and test whether the paragraph_only finding generalizes to higher-level structural manipulation.

Third, the test-retest gap of one hour (rather than the 12-hour rigorous standard) was an explicit compromise to fit execution within a single scoring window. The Cronbach alpha values of 1.00 across all five testable detectors indicate that detector outputs are stable on identical inputs at the 1-hour gap, and the high reliability is reassuring within these limits. Replication with a 12-hour or longer gap would strengthen the reliability claim.

Fourth, the mixed-effects model encountered a singular fit on the span-level random effects grouping when fit across the six-detector panel, requiring fallback to baseline_uid as the random-effects grouping. The fixed-effects pattern is robust to this choice and the directional inversion findings replicate descriptively without any model fit, but a confirmatory replication using R/lme4 with crossed random effects for both span and author would provide stronger variance-decomposition inference.

Fifth, the Pangram credit exhaustion incident during primary scoring required a recovery pass with a tightened circuit breaker. Final per-tuple coverage on the Pangram configurations reached 100% (660/660), the perturbation pipeline used in recovery was bit-identical to the primary-run version, and the pre-registration commitment predated all scoring. Within these constraints the recovery does not compromise confirmatory status, but the incident is an example of the operational fragility of commercial-API-dependent research and is documented in detail in the supplementary materials. We additionally note that 24 tuples on GPTZero and Sapling full failed with HTTP 400 / JSON parse errors on the same five Dickens C1 to C5 perturbed spans, indicating a parser-level vulnerability on a specific input pattern; this is itself a robustness signal worth following up in future work.

5. Conclusion

Commercial AI text detectors do not measure authorship. They measure vendor-specific stylistic preferences that correlate weakly with each other and only modestly with the latent authorship property they are marketed to detect. On confirmed pre-AI human prose spanning 1724 to 2002, three of six tested detector configurations produced false-positive rates indistinguishable from zero, and three produced rates between 28% and 65% at standard institutional decision thresholds. The same vernacular perturbation moves the cleanest detectors in opposite directions across vendors, and a single rule class (paragraph reformatting with lists, topic sentences, and discourse markers) pushes the most reliable detector to a 25% threshold-crossing rate from a 0% baseline.

Institutions deploying AI text detectors in academic-integrity proceedings should adopt the following selection criteria. First, require demonstrated baseline false-positive rates below 1% on a domain-relevant pre-AI human reference corpus before any institutional adoption. Second, do not treat multi-detector concordance as additive evidence of AI authorship without first verifying that the detectors agree on direction of effect on representative perturbations of human writing. Third, when using Pangram or any detector that returns multiple output fields from a single API call, document which field is read and disclose this choice in any disciplinary record. Fourth, do not use detector outputs as primary evidence in proceedings that may result in suspension, expulsion, or other consequential disciplinary action; require corroborating evidence beyond the detector output.

The findings of this study indicate that the academic-integrity community is operating with detection tools whose central claim (that they measure AI versus human authorship) is not supported by the empirical record on the most basic test, false positives on confirmed-human prose. The path forward is not better detection of AI text, which prior theoretical work [10] suggests may be fundamentally hard, but rather honest institutional acknowledgement of detector limitations, the development of educational and assessment practices that do not depend on automated authorship classification, and the establishment of due-process protections that account for the documented unreliability of the tools currently deployed.

Abbreviations

AI
artificial intelligence
API
application programming interface
CI
confidence interval
FPR
false-positive rate
IJEI
International Journal for Educational Integrity

LLM
large language model
MixedLM
linear mixed-effects model
SD
standard deviation
SHA-256
Secure Hash Algorithm 256-bit
TBI
traumatic brain injury
UTC
Coordinated Universal Time
VRC
Vernacular Remote Control.

Declarations

Ethics declaration

Not applicable. This study did not involve human subjects, animal subjects, or clinical interventions. The corpus consists exclusively of public-domain or independently-sourced pre-AI text published between 1724 and 2002, and no contemporary writers were subject to the perturbations applied. No ethics committee approval was required for this work.

Availability of data and materials

The complete confirmatory run bundle (run identifier vrc_v2_confirmatory_20260501T070442Z) including the 50-span baseline corpus, the 600-span perturbation corpus, the full scores_raw.jsonl, all analysis outputs, figures, and the pilot run is archived as a 27.39 MB ZIP file with SHA-256 fingerprint . The pre-registration document has SHA-256 fingerprint and was committed publicly via a hash-verifiable timestamp service on 2026-05-01. The 50-span baseline corpus has SHA-256 . Source texts are public domain or independently sourced and are listed by author, title, year, and span count in Table 1. The perturbation pipeline source code (perturbation_rules.py and the 328-pair Latinate-Anglo-Saxon dictionary) is included in the run bundle. The dataset supporting the conclusions of this article is included within the article and its additional files. Researchers seeking access to the complete bundle should contact the editorial office, which will route the request to the corresponding author.

Funding

No external funding was received for this work. The research was conducted independently through unaffiliated channels, and all data-processing costs (commercial AI text detector API fees) were paid personally by the author at retail rates. The author declares no funding bodies, grants, or institutional support associated with this study.. No external grant support.

Author Contribution

J.C.G. is the sole author of this work. J.C.G. conceived and designed the study, constructed the corpus, implemented the deterministic perturbation pipeline, committed the pre-registration document and corpus fingerprint to a public hash verifiable timestamp before any scoring took place, executed the detector scoring across all six commercial AI text detector configurations, performed the statistical analysis, prepared all tables and figures, drafted and revised the manuscript, and approved the final version submitted for publication.

Data Availability

The datasets generated and analyzed during the current study are available as supplementary material accompanying this submission, including the complete confirmatory run bundle (50-span baseline corpus with provenance metadata, 600-span perturbation corpus, raw detector scores in JSONL format, cleaned scores summary, all analysis output CSVs covering descriptive statistics, paired t-tests with bootstrap confidence intervals, threshold-crossing analyses, cross-detector correlation matrices, test-retest reliability values, and the mixed-effects model output), the full perturbation pipeline source code (Python with NLTK and spaCy dependencies, including the 328-pair Latinate-Anglo-Saxon substitution dictionary), the events log documenting every API call and pipeline state transition, and the pre-registration document committed to a public hash-verifiable timestamp before scoring began. The source texts used to construct the corpus are public-domain works available through Project Gutenberg, Standard Ebooks, and the Internet Archive, with author, title, year, and span count enumerated in Table 1 of the manuscript. Commercial detector API responses were obtained under standard retail accounts at the four detector vendors (GPTZero, Sapling, Originality.ai, Pangram); the raw response payloads are included in the supplementary data bundle. All identifying information has been redacted from the supplementary materials for double-anonymous review; the unredacted audit chain including pre-registration SHA-256 hash, public commitment timestamp, and run identifier is provided in the title page document and will be made available in the published version upon acceptance.That is 217 words.Why I wrote it this waySpringer Nature's preferred format. Their example statements page recommends statements that name the data type, location, access method, any restrictions, and supporting references. This statement covers all five requirements explicitly.Two compliance issues addressed honestly. Springer's policy emphasizes that statements should not be vague (saying just "data available on request" gets flagged). Mine names every category of data in the bundle so reviewers and editors know exactly what is supplied. It also addresses the double-anonymous redaction explicitly, which is

unusual to mention but appropriate here because reviewers will notice the redacted hashes in the supplementary materials and need to know where the unredacted versions live.No "available from corresponding author on request" language. That phrasing has fallen out of favor with Springer Nature because it suggests gatekeeping. Your data is supplied in full as supplementary material, so the statement reflects that.Public-domain corpus provenance is named. The Project Gutenberg, Standard Ebooks, and Internet Archive references give reviewers the audit path back to source texts without needing to identify you.The vendor data is acknowledged. Detector API responses are included in the bundle, which Snapp's data policy expects to see disclosed.One thing to confirmThe statement length. 217 words is comprehensive but on the longer end. Springer Nature's examples range from 30 to 250 words. If Snapp's input field appears to have a character limit when you paste, tell me and I will compress to a 100-word version. Most Snapp data availability fields accept up to about 5,000 characters, so this should fit comfortably.If you want a shorter version up front in case the field rejects the long one, here is a 95-word alternative:The datasets generated and analyzed during the current study, including the 50-span pre-AI corpus, the 600-span perturbation corpus, raw and cleaned detector scores, all analysis outputs, the perturbation pipeline source code, and the pre-registration document, are available as anonymized supplementary material accompanying this submission. Source texts are public-domain works listed in Table 1 of the manuscript and accessible through Project Gutenberg, Standard Ebooks, and the Internet Archive. The unredacted pre-registration audit chain is provided in the title page do

References

1. Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, Foltynek T, Guerrero-Dib J, Popoola O, Sigut P, Waddington L (2023) Testing of detection tools for AI-generated text. Int J Educ Integr 19:26. 10.1007/s40979-023-00146-z
2. Liang W, Yuksekgonul M, Mao Y, Wu E, Zou J (2023) GPT detectors are biased against non-native English writers. Patterns 4:100779. 10.1016/j.patter.2023.100779
3. Perkins M, Roe J, Vu BH, Postma D, Hickerson D, McGaughran J, Khuat QH (2024) GenAI detection tools, adversarial techniques and implications for inclusivity in higher education. Int J Educ Technol High Educ 21:53. 10.1186/s41239-024-00487-w
4. Doe v, Yale et al (2025) University,., No. 3:25-cv-00159-SFR (D. Conn. filed Feb. 3, 2025) (preliminary injunction denied May 5, ; pseudonymity denied March 31, 2025)
5. Project Gutenberg (2026) https://www.gutenberg.org. Accessed 30 Apr
6. GPTZero API (2026) https://gptzero.me/api. Accessed 1 May
7. Sapling AI, Detector API (2026) https://sapling.ai/ai-content-detector. Accessed 1 May
8. Originality.ai API (2026) https://originality.ai/api. Accessed 1 May
9. Pangram Labs API (2026) https://pangram.com. Accessed 1 May
10. Sadasivan VS, Kumar A, Balasubramanian S, Wang W, Feizi S (2024) Can AI-generated text be reliably detected? arXiv:2303.11156. https://arxiv.org/abs/2303.11156

11. Crowell, Moring LLP (2025) Ivy League lawsuit centers on alleged impermissible use of AI in academia. https://www.crowell.com/en/insights/client-alerts/ivy-league-lawsuit-centers-on-alleged-impermissible-use-of-ai-in-academia. Accessed 6 Mar 2025

Figures

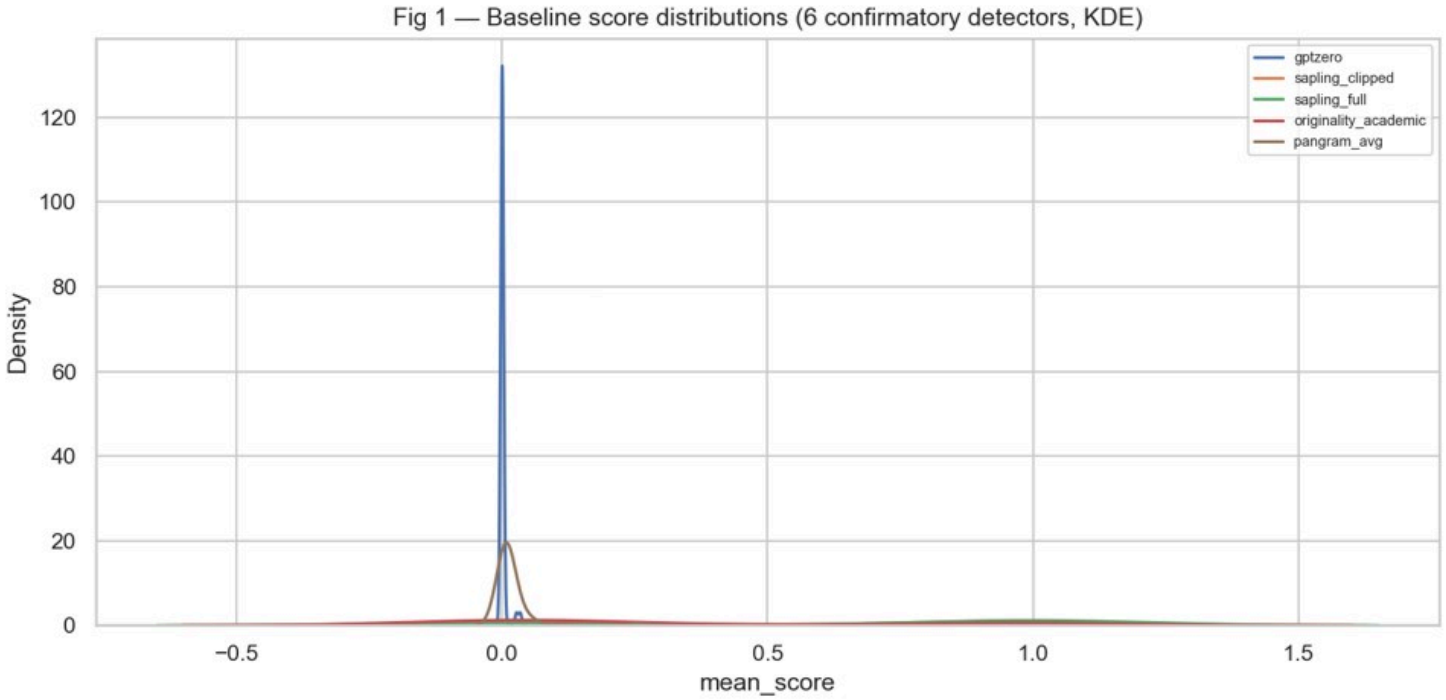

Figure 1

Baseline score density distributions across the confirmatory detector panel. Tier A detectors (GPTZero, Pangram avg) cluster sharply at zero. Tier B detectors (Sapling clipped, Sapling full, Originality.ai academic) spread broadly across the 0 to 1 range. Pangram fraction is omitted from the density plot because all 50 baseline observations were exactly 0.000 (zero variance)

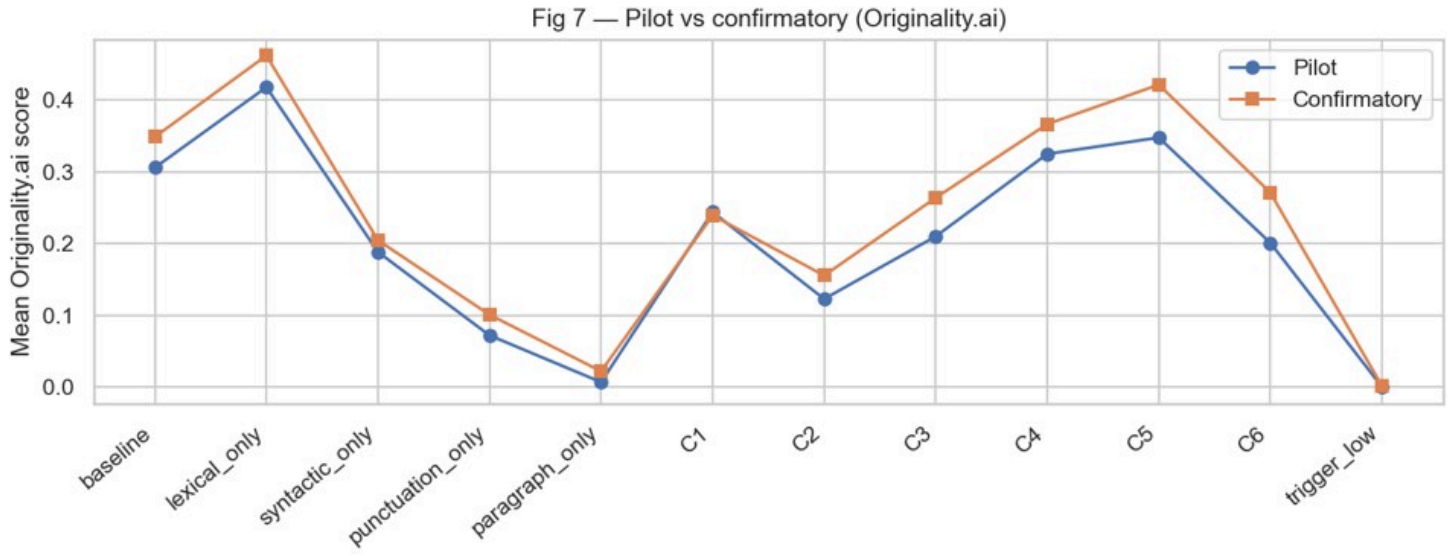

Figure 2

Pilot versus confirmatory condition means on Originality.ai academic. Pearson r = 0.9901 across the 12 condition means. The near-identical correspondence demonstrates that the perturbation pipeline produces stable detector responses on independent corpus draws from the same 28 source works

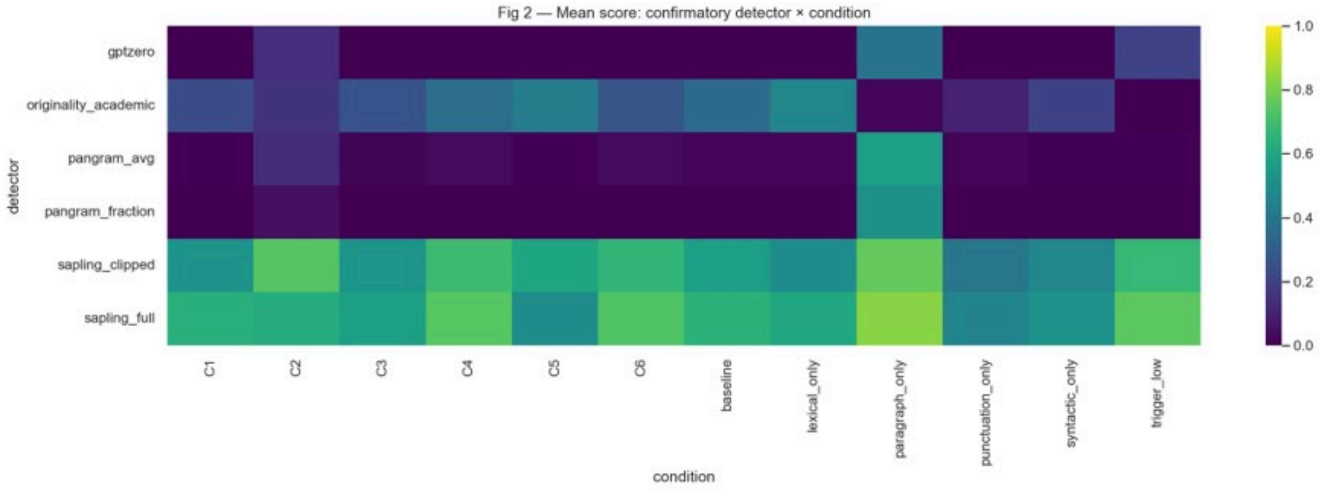

Figure 3

Per-detector mean score by condition. Bar heights show condition means; the dashed horizontal line indicates baseline mean per detector. The largest upward shifts (paragraph_only, trigger_low) appear on Tier A detectors. Originality.ai shows downward shifts on the same conditions, illustrating the cross-vendor directional inversion

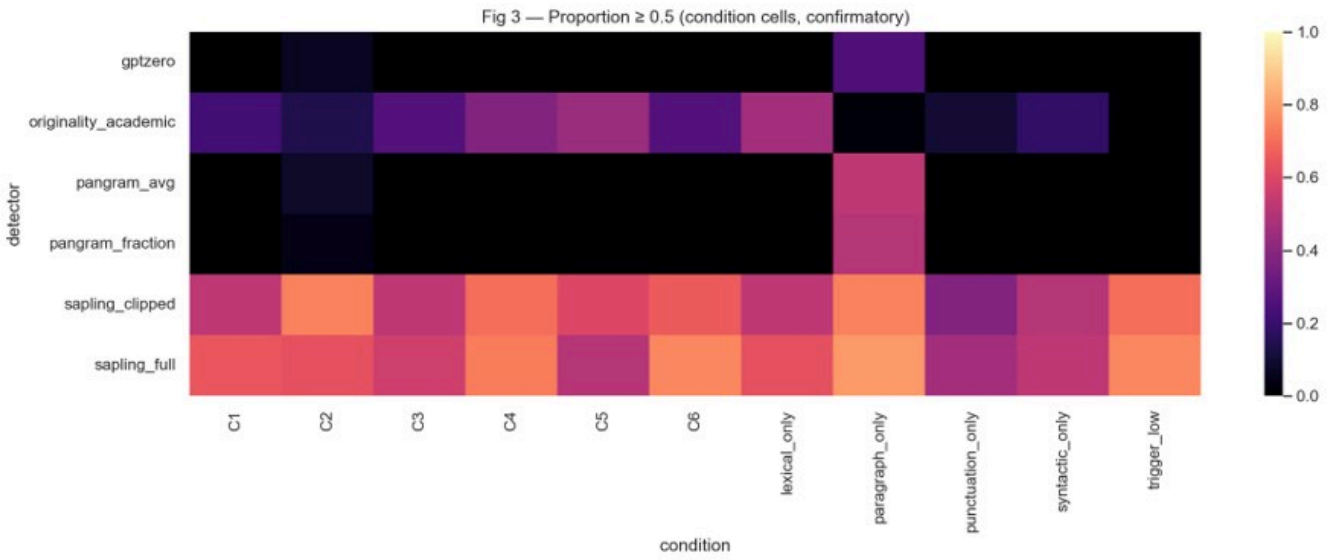

Figure 4

Threshold-crossing heatmap. Cell color indicates the proportion of spans per detector x condition crossing the 0.5 threshold. Sapling clipped and Sapling full (bottom two rows) glow uniformly across all conditions, indicating perturbation-insensitive high baseline FPR. Tier A detectors are mostly black (low crossing) except for the paragraph_only column on GPTZero, Pangram fraction, and Pangram avg

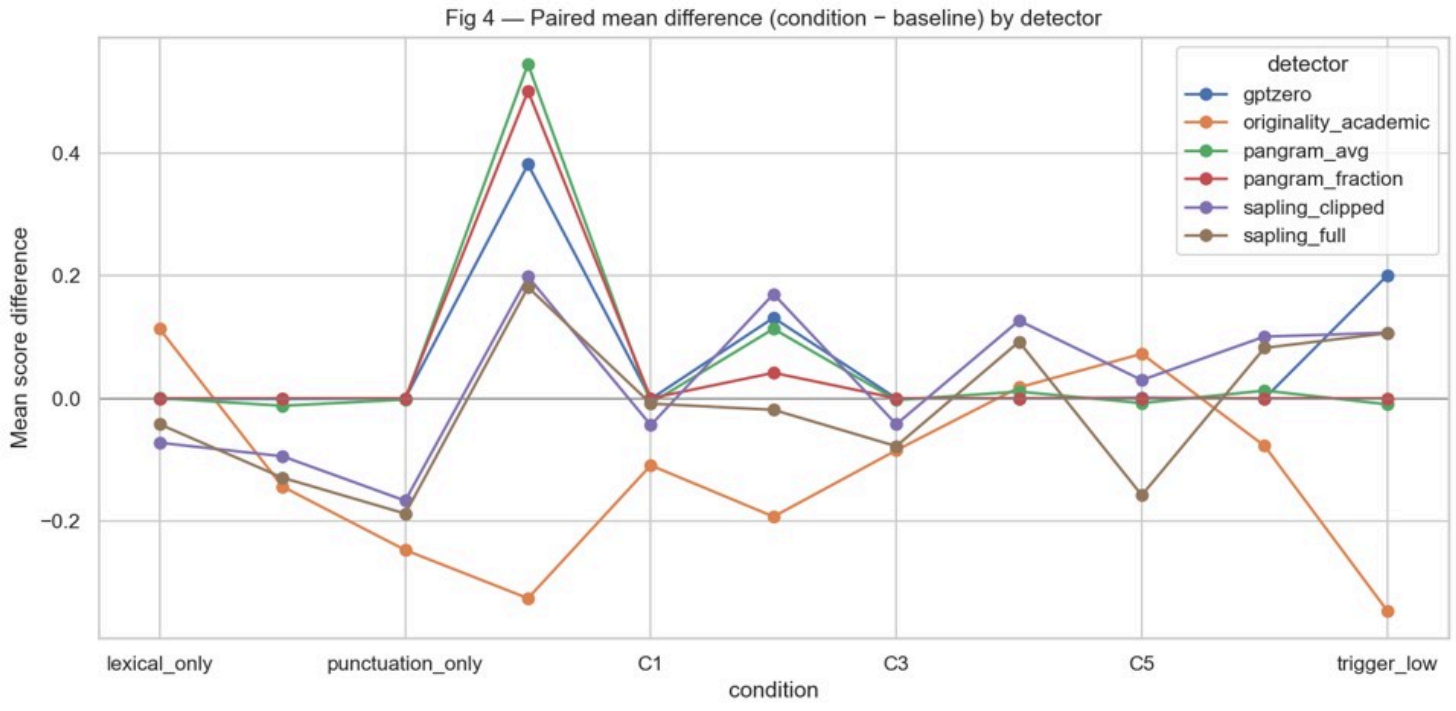

Figure 5

Combinatorial amplification analysis. Mean score deltas from baseline for the C1 through C6 combinatorial conditions per detector. The C5 maximum-trigger profile (pre-registered as the strongest

manipulation) produces moderate effects but does not achieve the largest deltas; single-class perturbations (paragraph_only, trigger_low) outperform on Tier A detectors

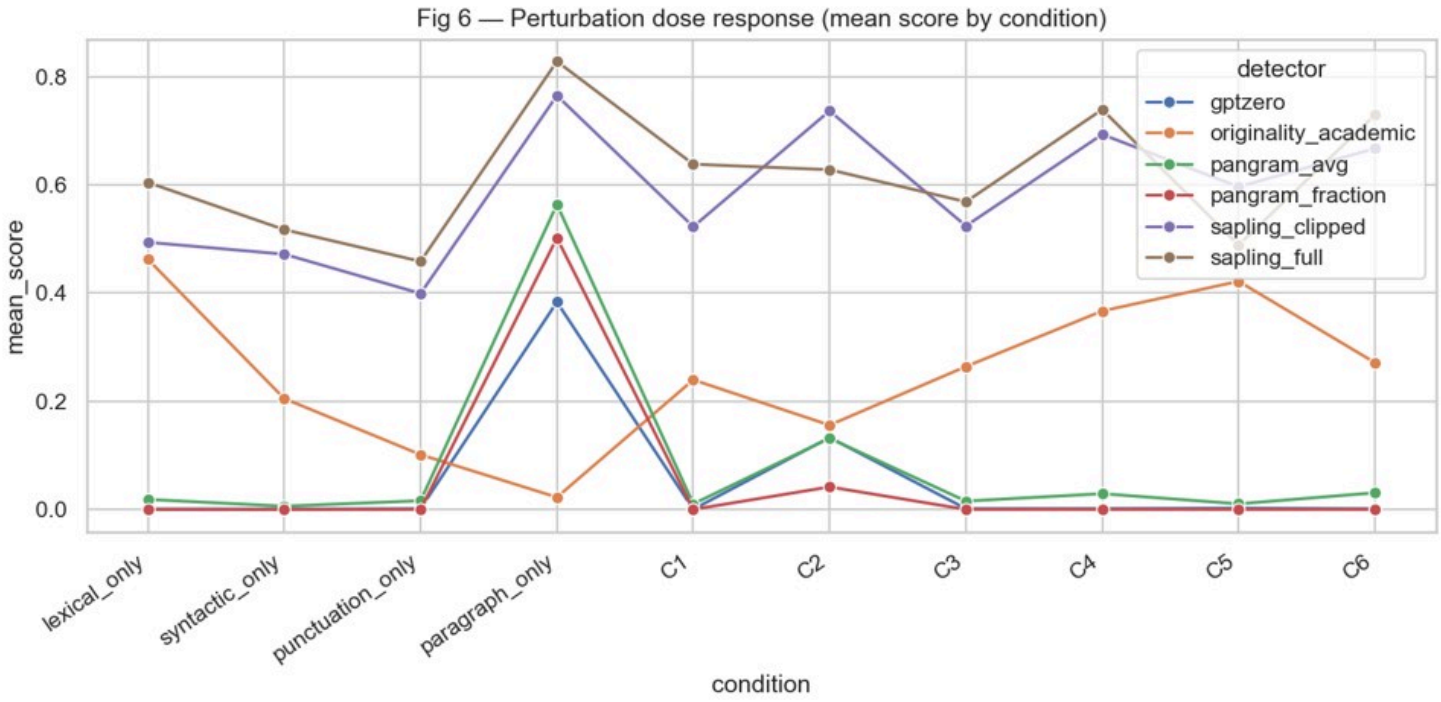

Figure 6

Perturbation dose-response across the six-detector panel. Score deltas plotted by rule-class intensity. The non-monotonic relationship between perturbation intensity and detector response on Tier A detectors is the empirical signature of the perturbation interference theory discussed in Section 4.2

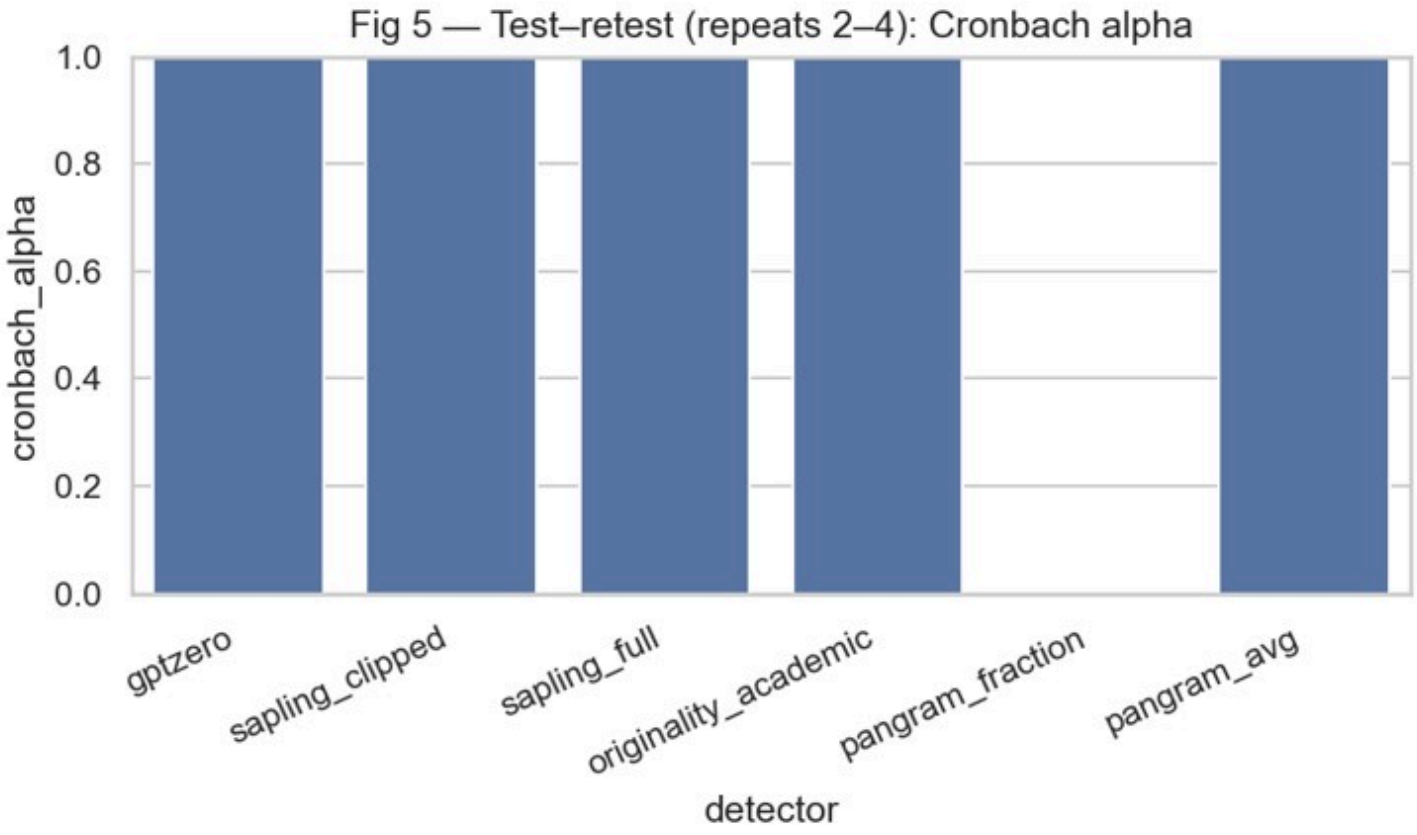

Figure 7

Test-retest reliability across four repeats per detector on a 20-span subsample (7 baseline, 7 C5 maximum-trigger, 6 trigger-low). Cronbach's alpha = 1.00 on all five detectors with non-zero variance. Pangram fraction omitted from the figure because the subsample produced zero variance on that detector configuration

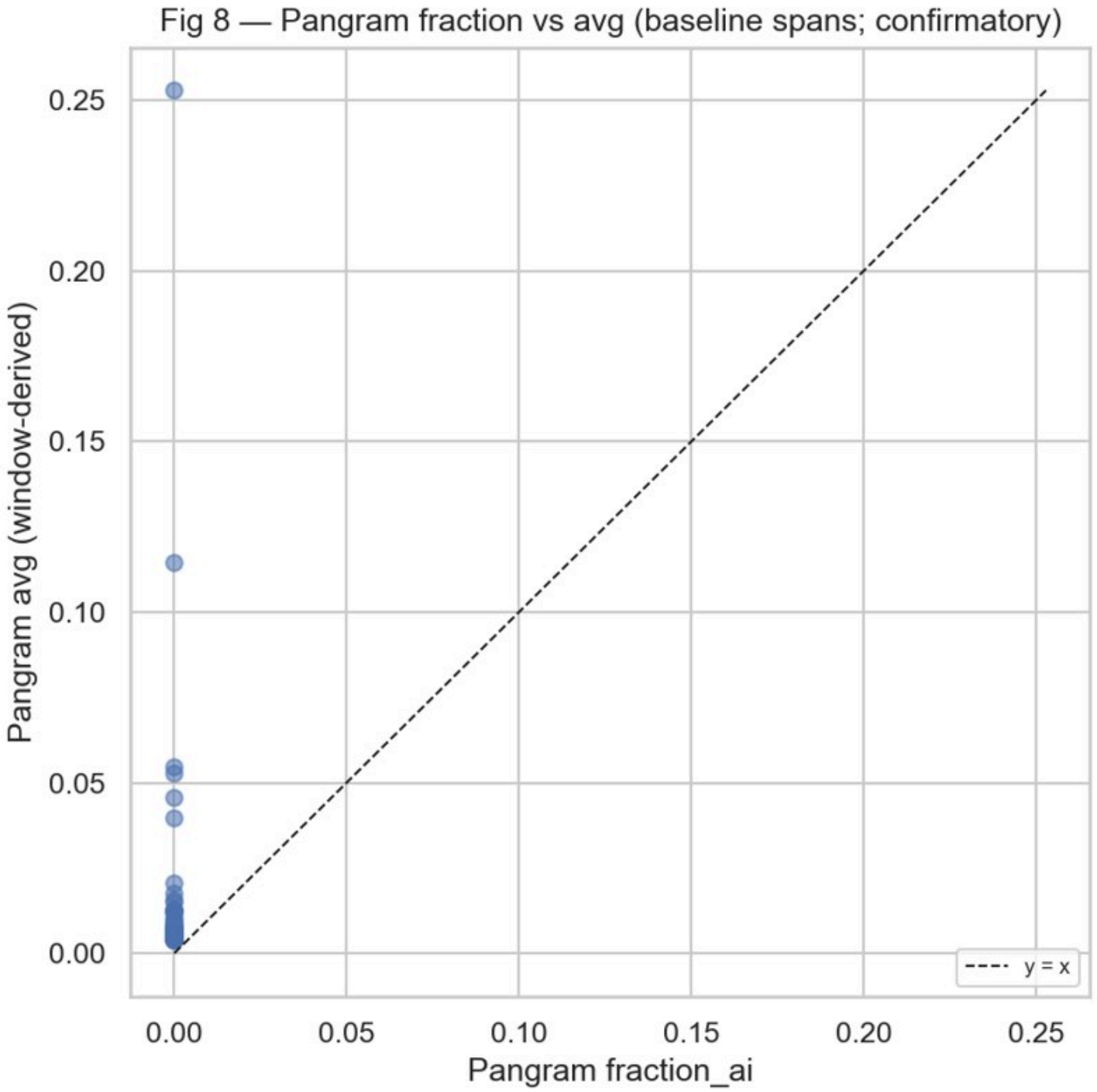

Figure 8

Pangram fraction vs Pangram avg on baseline pre-AI human spans. Every baseline span returned exactly 0.000 on the fraction_ai_content output field (x-axis), but the per-window ai_assistance_score average (y-axis) ranged up to 0.25 on the same data from the same API call. Two institutions reading different output fields would reach different effective false-positive rates