A Controlled Comparison of AI Text Humanizers on Academic Writing

TextPulse Research. Working Paper.

Abstract

AI text humanizers rewrite machine-generated text so that AI detectors classify it as human. Humanizer vendors advertise pass rates, but no published study measures what the rewriting does to the output text itself. We passed 48 AI-generated academic texts (about 300 words each, four disciplines, three citation styles, four AI model generator families) through eleven humanizers under one setting per tool, and scored 432 outputs on a fixed set of metrics, namely, position on a stylometric human-to-AI spectrum; semantic faithfulness to the source; retention of technical terms, numbers and citations; fabricated numbers and named entities; Flesch-Kincaid grade shift; grammar errors introduced; length change; rewrite depth; hidden-character injection; and vocabulary shift against a 1,057-word AI vocabulary lexicon generated from our prior study. Detection was measured with Turnitin and GPTZero, the gold standard university-grade detectors. Every humanizer tool pushed the text toward the human end of the spectrum, reduced AI-leaning vocabulary and lowered the reading grade, and the tools differed most in what they preserved. TextPulse rewrote 52 percent of the words while keeping 97 percent of the document meaning, 99 percent of citations and 87 percent of numbers, introduced 0.65 grammar errors per 1,000 words, fabricated almost no named entities, and removed more AI-leaning vocabulary than any tool that kept the text intact. The tools with the highest raw GPTZero pass rates had the weakest content: Walter Writes passed 60 percent of its outputs but also fabricated 4.4 numbers or names per each (~300 word) text, and Phrasly passed 38 percent but introduced 3.6 grammar errors per 1,000 words. On a strict experiment, when a detector pass is counted only for outputs that also preserve meaning, citations and facts, TextPulse has the highest rate at 10.4 percent (95 percent CI 2.1 to 18.8), Phrasly and GPTinf have 2.1 percent, and eight tools have none. TextPulse is also first on an equal-weight composite of rewrite depth, faithfulness, spectrum position, grammar cleanliness and content integrity, with or without the detector component. On Turnitin, a document holding only TextPulse's outputs scored 24 percent AI, the lowest score in the field together with StealthWriter; the other tools scored from 28 to 74 percent. Detector pass rate on its own is an incomplete measure of a humanizer, since the easiest approach to pass is to change what the text says, or even intentionally inject grammar errors. Across all experiments, all of the humanizer tools had the same treatment to allow for an even level playing field. Every metric was scripted, and all humanization outputs and code are made publicly available for easy replication of the same experiments in future work.

1. Introduction

A humanizer can be considered as a rewriting tool. It takes text written by an AI language model and outputs a “humanized” version that AI detectors are less likely to flag. The market grew quickly after universities adopted detection software, and the tools now advertise their results as pass rates against popular detectors such as Turnitin, GPTZero and Originality AI. However, the advertising does not report the condition of the text after rewriting. An academic text typically carries numbers, citations, technical terms, and a formal, academic register. A rewrite that drops or fabricates any of these, or turns a formal paragraph into a casual one, has not preserved the text, regardles of whether it passes a detector or not.

This study measures both sides simultaneously. Eleven top-tier humanizers rewrote the same 48 AI-generated academic texts. Each output was scored for detection, for the degree the original content survived, for fabrication rate, and for change in style. Since each tool received identical inputs, differences between outputs can be considered differences between the humanizer tools.

Earlier papers built the measurement tools reused here, namely, a stylometric feature set (TextPulse Research, 2026e), a human vs. AI classifier and the spectrum derived from it (2026f), a lexicon of AI-preferred vocabulary (2026d), a citation audit method (2026c), and a comparison of commercial AI detectors (2026a). The conflict of interest is managed by protocol. Every tool received the same inputs, every metric is a public script, and every raw output was released.

This study aims to answer four key questions. Where does each tool’s output fall on the human-to-AI spectrum? How much of the source meaning, facts and citations does each tool preserve, and what does it fabricate? Which tools pass a commercial detector, and does passing come at the cost of ‘damaging’ the text? What does the trade-off between passing and preserving look like across the humanizer tools?

2. Related work

2.1 How AI text detectors work

Detection research began with the statistics of the generator itself. GLTR overlaid token rank and entropy on text and increased human detection of GPT-2 output from 54 to 72 percent without any training (Gehrmann et al., 2019). Solaiman et al. (2019) fine-tuned a RoBERTa classifier that identified GPT-2 output at about 95 percent accuracy, the supervised baseline that later commercial detectors followed. Ippolito et al. (2020) showed that the decoding choices that fool human readers leave statistical traces that make automatic detection easier, and Uchendu et al. (2020) framed the problem as authorship attribution. Fröhling and Zubiaga (2021) showed that hand-built stylometric features match expensive neural detectors, and Desaire et al. (2023) separated chemistry papers from ChatGPT versions with 20 such features at over 99 percent document-level accuracy. The stylometric spectrum used in this study belongs to that family (TextPulse Research, 2026e, 2026f).

Zero-shot methods read the generator's probabilities. DetectGPT uses the curvature of log-probability around a text (Mitchell et al., 2023); Fast-DetectGPT replaces its perturbation step with sampling and runs about 340 times faster (Bao et al., 2024); Binoculars contrasts the perplexity of two related models and detects over 90 percent of ChatGPT samples at a 0.01 percent false positive rate (Hans et al., 2024). Perplexity is the standard measure of the 'predictability' of a text under a language model, and low perplexity is treated as a machine signal. Burstiness, or variation in sentence length, is the second signal in GPTZero's public description. Tulchinskii et al. (2023) found that the intrinsic dimension of human text embeddings is about 1.5 higher than that of its AI counterpart. Supervised detectors have been trained against attacks. RADAR trains a paraphraser against the detector (Hu et al., 2023), OUTFOX uses adversarially generated examples in context (Koike et al., 2024), Ghostbuster searches over weak-model features and reaches 99.0 F1 (Verma et al., 2024), and Pangram uses hard-negative mining (Emi & Spero, 2024; Glickenhaus et al., 2026). Watermarking embeds a detectable signal during generation (Kirchenbauer et al., 2023). Surveys by Jawahar et al. (2020), Crothers et al. (2023), Tang et al. (2024), Yang et al. (2024) and Wu et al. (2025) agree that paraphrasing, distribution shift as well as mixed human-AI text are the open problems.

On RAID, over six million generations across 11 models and 11 adversarial attacks, detectors were fooled by repetition penalties, sampling changes and homoglyphs (Dugan et al., 2024). MAGE found that the best detector identified only 84 percent of out-of-domain texts from an unseen model (Li et al., 2024), M4GT-Bench and MGTBench found that good performance generally requires training data from the same domain and also the same generator (Wang, Mansurov et al., 2024; He et al., 2024). Doughman et al. (2025) found performance being reduced to chance on easy-to-read texts. Beemo showed that expert human editing of model output evades detection while model-edited text rarely does (Artemova et al., 2025). ARB (Perrone & Romano, 2026) and Baidya et al. (2026) both report that heavy humanization is the dominant failure mode of current AI detectors.

2.2 Evasion by paraphrasing and humanization

Paraphrasing passes detectors when at scale. DIPPER, an 11-billion-parameter paraphraser, decreased DetectGPT's detection rate from 70 to 5 percent at a 1 percent false positive rate and evaded watermarking, GPTZero and OpenAI's classifier (Krishna et al., 2023). Sadasivan et al. (2025) showed that recursive paraphrasing reduced watermark detection from 99.8 to 9.7 percent and tied the best possible detector performance to the distance between human and AI text distributions. Shi et al. (2024) passed watermarking, GPTZero and DetectGPT with word substitution and instructional prompts, Zhou et al. (2024) changed detector verdicts with small perturbations in about 10 seconds, RAFT compromised every tested detector with word-level substitutions that human raters could not distinguish from the originals (Wang, Li et al., 2024), and Jovanović et al. (2024) scrubbed and spoofed watermarks with a small set of cheap API queries.

Cheng et al. (2025) reduced true positive rates by an average of 88 percent with a detector-guided paraphrase. AuthorMist trained a 3-billion-parameter model with reinforcement learning against GPTZero, Winston, Originality.ai and Sapling and reached 79 to 96 percent attack success while maintaining a semantic similarity score over 0.94 (David & Gervais, 2025). MASH reached 92 percent success by injecting human style (Gu et al., 2026), StealthRL trained against an ensemble of detectors (Ranganath & Ramesh, 2026), Pedrotti et al. (2025) shifted a model's writing style to clear detectors trained on its original output, and Xu et al. (2026) found that raw base-model output already reads as human to GPTZero and Pangram. TH-Bench evaluated six humanization approaches against thirteen detectors across nineteen domains and found that no approach wins on evasion, text quality and cost at once (Zheng et al., 2025). Bao et al. (2025) built a risk-level benchmark that ends in a humanization tier. DAMAGE audited nineteen commercial humanizers and paraphrasers and found that most existing detectors miss their output (Masrour et al., 2025). Russell et al. (2025) found that those who write with ChatGPT preserve their accuracy on humanized text where most automated detectors lost theirs. *Each of these studies measures whether a detector is passed, but none of them scores the quality of the humanized text itself.*

2.3 Commercial detectors in education

Evaluations of commercial detectors on academic text reach the same conclusion from different corpora. Weber-Wulff et al. (2023) tested fourteen tools and found all below 80 percent accuracy, with machine paraphrasing and manual editing lowering accuracy further and about a fifth of AI texts attributed to humans. Walters (2023) found that only Copyleaks, Turnitin and Originality.ai were accurate on GPT-3.5, GPT-4 and student essays. Elkhatat et al. (2023) reported lower accuracy on GPT-4 than GPT-3.5 output and false positives on human engineering text. Chaka (2023, 2024a, 2024b) found no tool consistently reliable across chatbots or across 30 detectors and concluded from the literature that detectors are unreliable for authorship decisions. Gao et al. (2023) found that blinded reviewers identified only 68 percent of generated abstracts. Ibrahim, Liu et al. (2023) found that ChatGPT matched students in 9 of 32 courses and that classifiers misclassified a substantial share of its answers. Perkins, Roe, Postma et al. (2024) found that Turnitin flagged 91 percent of prompt-engineered GPT-4 submissions but identified only 55 percent of their content, and Perkins, Roe, Vu et al. (2024) reduced detector accuracy by 17 percentage points with simple techniques such as misspellings and added burstiness. Ibrahim, Al Otaibi and Sibai (2025) found Turnitin's detector close to chance once paraphrasing tools were applied, Kar et al. (2025) found that free detectors varied widely on paraphrased text, and Barrot and Aranda (2025) found detectors failing to correctly classify AI-edited essays.

Recent 2025 and 2026 studies address humanizers directly. Jabarian and Imas (2025) found that of Pangram, Originality.ai, GPTZero and RoBERTa only Pangram held a 0.5 percent false positive rate on humanizer output. Van Vlasselaer et al. (2026) placed 37 of 40 humanized papers in the expected AI range with Pangram, exceeding Turnitin, GPTZero and Copyleaks. Hadra et al. (2026) found Originality.ai at 0.69 and Turnitin at 0.51 overall accuracy with hybrid-text recall falling to

0.31 and 0.02. Sun et al. (2026) found that hybrid editing reached 88 percent of AI content evade detection, and Karr et al. (2026) found that light edits to an abstract were flagged at 64 to 80 percent while unmodified recent abstracts were flagged at 9 to 15 percent, with scores tracking long-word density. Turnitin announced a bypasser-detection feature in August 2025 without published accuracy figures (Turnitin, 2025; Bailey, 2025). Policy commentary argues that probabilistic detector scores cannot meet the standard of proof used in misconduct cases (Bassett et al., 2026; Zhang & Lv, 2026; Ardito, 2025). Our own inter-rater study found nine commercial detectors agreeing on pure human and pure AI text and disagreeing at chance level on mixed human-AI text (TextPulse Research, 2026a).

2.4 False positives

Liang et al. (2023) reported that seven detectors flagged 61 percent of TOEFL essays by non-native writers as AI while US eighth-grade essays were almost never flagged. Our replication with a transparent stylometric classifier found no bias against non-native writers for that classifier, and a threefold bias for a gradient boosting model trained on the same features, so the bias is a property of the detector and not of the writing itself (TextPulse Research, 2026g). Garland (2026) shows that population-level variation in style guarantees disproportionate false positives for some groups, and Hu (2026) mentions that fixed lexical markers push scholars to flatten their prose. Any pass rate reported for a humanizer needs the detector's false positive rate on human text with it, which this study provides.

2.5 What rewriting does to vocabulary and style

Language models change vocabulary in a recognizable direction. Kobak et al. (2025) measured excess vocabulary across 15 million biomedical abstracts and estimated that at least 13.5 percent of 2024 abstracts were processed with a model. Liang, Izzo et al. (2024) estimated that 6.5 to 17 percent of peer-review text at AI conferences was substantially model-modified, and Liang, Zhang et al. (2024, 2025) mapped the increase to 22 percent of computer science papers. Matsui (2024), Astarita et al. (2024) and Geng and Trotta (2024) found the same words overused in PubMed, astronomy and arXiv abstracts, Yakura et al. (2025) found them in spoken academic talks, and Juzek and Ward (2025) traced the overuse to reinforcement learning from human feedback (RLHF) tuning. Sourati et al. (2026) found that model polishing reduces the variance of writing complexity by 21 to 50 percent across 880,000 texts.

In a paired corpus of 60,786 human texts and their AI rewrites, the AI-leaning vocabulary was 44 percent Latinate against 10 percent for human-leaning words, and the strongest markers were formal connectives and substitutions such as "thereby", "consequently" and "utilized" (TextPulse Research, 2026d). The same corpus showed that rewriting alters or drops citations at measurable rates (2026c) and changes punctuation, sentence-length variation and syntax in ways that identify the AI model (2026e). A humanizer is a rewriter aimed in the opposite direction, from machine style toward human style, and the same measurements apply.

2.6 The gap

In short, the literature on detector evasion measures whether the detector is bypassed. The education literature measures whether a detector catches humanized work. Neither holds inputs constant across top-tier commercial humanizers and scores what each tool does to meaning, citations, numbers, entities, grammar and register of the humanized text. The related studies are DAMAGE's qualitative audit of nineteen tools (Masrour et al., 2025), TH-Bench's six attack methods (Zheng et al., 2025), Van Vlasselaer et al.'s single humanizer path (2026) and Jabarian and Imas's four detectors on humanizer output (2025). Vendor and affiliate comparisons report pass rates against one or two detectors on a handful of texts. HumanizerBench, the most cited leaderboard, is operated by WriteHuman and ranks WriteHuman first. This study fills that gap with thirteen tools, identical inputs, and content scoring beside detection.

3. Data and methods

3.1 Source texts

The input set is 48 generated Results and Discussion texts, 12 per each of four language models, through their APIs, namely, OpenAI gpt-5.6-terra, Google gemini-3.7-flash, DeepSeek deepseek-v4-pro, and Anthropic claude-sonnet-5. Reasoning was disabled on every provider that supports it, since how a generator thinks before writing a fictional text has no affect on a comparison between humanizers. Each text covers one of 48 invented empirical studies across four primary disciplines (medicine and health, physical and computational sciences, social sciences, humanities) in one of three citation modes (APA author-year, numeric brackets, or citation-free). Texts were at an average length of 301 words (range 264 to 344).

Each generator was instructed to list, after the text, the technical terms, every numeric fact, and every in-text citation it used. These lists were verified against the text, and any listed item not present verbatim in the text was dropped. The verified ground truth holds 404 terms, 887 numeric facts and 124 citations. Ground truth was therefore never a model's self-report, but a manual check.

3.2 Tools and runs

Eleven tools were run, namely, TextPulse, Grammarly AI humanizer, QuillBot AI Humanizer, Humanize AI Pro, StealthWriter, Walter Writes, HIX Bypass, GPTinf, UnAIMyText, Phrasly, and Clever Humanizer. WriteHuman and Undetectable.ai were also attempted and then excluded from the study: WriteHuman caps its output at 250 words, which is too short to carry the samples, and Undetectable.ai returns its own prompt preamble, such as "Here is the rewritten text", inside the output text. All humanizations took place between 26 and 28 August 2026. The user pasted each text into each tool's web interface once, under the tool's default setting, and pasted the output back verbatim through a console that recorded the original AI input and humanized output pair. No output was edited. AI humanizer tools were identified by product name throughout.

Free tiers limit how many texts a tool can process, so per-tool n is unequal by design. A priority core of 24 texts, the same for every tool and in the same order, was defined so that the paired comparison holds at whatever depth a tool reached. Final counts are as follows: TextPulse 48, Walter Writes 48, GPTinf 48, UnAIMyText 48, Grammarly 47, Humanize AI Pro 47, Phrasly 47, StealthWriter 34, Clever Humanizer 24, QuillBot 22, HIX Bypass 19, amounting to a total of 432 scored outputs. Every summary reports its own n.

Truncation by free tiers was kept as data. QuillBot returned about 125 words per text and GPTinf capped at about 300 words. No output was edited.

3.3 Metrics

All metrics were computed by publicly available scripts from the source text and the output. Missing values are null.

Stylometric spectrum

Each output is placed on a line where the median human academic text is 0 and the median AI text is 100. The line is built from 47 stylometric features measured in earlier work (TextPulse Research, 2026e) on 121,092 academic texts (2026f): sentence-length mean, spread and coefficient of variation (burstiness), word length, lexical diversity, punctuation rates, passive constructions, nominalisations, first-person pronouns, sentence-opener repetition, connective openers, contractions, -ly adverbs, long-word rate, and 25 AI-marker phrases. The features are standardised, and the axis is the direction from the human centroid to the AI centroid. The frozen model reproduces the earlier paper's out-of-fold AUC of 0.936, and scoring 250 human student essays as a check placed their median at -93. The middle half of the human anchors is from -36 to 33 on the axis, and a tool whose mean is inside that range is reported as reaching the human range. Tools inside the range are listed by sample size since their positions are statistically tied. The features that carry the most weight are mean word length and the rate of long words (both toward AI), lexical diversity, nominalisations and -ly adverbs (toward AI), and sentence-length burstiness and sentence-opener repetition (toward human). Section 4 gives an example of each one.

Semantic faithfulness

Cosine similarity between BGE-large embeddings of the whole source and the whole output, plus a sentence-alignment score and a count of output sentences with no source counterpart. In calibration, a faithful paraphrase scored 0.969 and a text with its second half deleted scored 0.640.

Named-entity and number retention

Named-entity retention is the share of the source's named entities (people, organisations, places, instruments, datasets, found in the source by a local named-entity model) that survive in the output. An entity is kept when every content word of its name appears in the output on a light stem, so inflection, word order and reformatting such as "Okafor & Linden" for "Okafor and Linden" all match. Generic vocabulary is never counted. Number retention is the share of verified numeric

facts that appear in the output. A number is counted as corrupted when its surrounding context survived but the value changed. This is a lower bound, because a rewritten sentence yields “dropped”, not “altered”.

Citation integrity

Citations are extracted from both texts and matched on unique author-year keys or bracket numbers. Presentation changes never count against a tool, i.e., a parenthetical citation rewritten as a narrative one, a semicolon list split into separate citations, a merged bracket such as [4,5] for [4] and [5], or a range [4-6] all match. A changed year or surname is considered a miss. A citation with no source counterpart is considered fabricated.

Register

Flesch-Kincaid grade of the output minus the source. AI Model-generated academic text reads at a higher grade than published human academic prose, and so a decrease moves the text toward the simplified formal style of human academic writing and away from the high-grade style of model output. A decrease is the desirable direction for humanized text.

Grammar

LanguageTool (v 6.8) errors per 1,000 words in the output minus the source, counting only rule categories that mark critical errors, but not style.

Length zones

Output-to-source word ratio in two zones: within tolerance, from 0.70 to 1.30 times the source length, and outside tolerance beyond that. Tools are reported in three length categories by the share of outputs within tolerance: kept length (95 percent or more), mostly kept (75 to 95 percent) and often broke length (under 75 percent). Within a category tools are listed by sample size.

Rewrite depth

The amount of words changed compared to original text, under a word-level diff, and the share of source word trigrams that survived verbatim.

Fabrication

Numbers in the output that do not occur in the source (years excluded), plus named entities (people, organisations, places, instruments) recognized by a local named-entity model none of whose content words occur in the source. Reformatted names such as “Okafor and Linden” for “Okafor & Linden” do not count.

Hidden-character injection

Characters absent from the source that are invisible, non-standard spaces, or letters from the Cyrillic or Greek blocks, with isolated Greek letters treated as scientific notation. (Some humanizers intentionally inject these into the output as a detection bypass attempt.)

Vocabulary shift

Hits per 1,000 words against the a 1,057-word AI vocabulary lexicon (TextPulse Research, 2026d), reported as the change in AI-favored words (581 entries) and as the change in an index that subtracts human-leaning hits (476 entries) from AI-leaning hits. The lexicon was frozen for the sake of this study.

3.4 Detectors

Turnitin

Each tool's outputs were combined into one document holding only that tool's outputs, one text per page, and submitted to Turnitin once. Turnitin returns one document-level AI percentage, the share of the document it attributes to AI writing, and prints an asterisk instead of a value under 20 percent. That document score is the Turnitin measure used here. An earlier submission that packed several conditions into shared documents, one of which mixed untreated AI texts with humanized outputs, was discarded, because the score of a mixed document cannot be attributed to any one condition.

GPTZero

Every text was sent once to GPTZero's API (model version 2026-08-09-base) and the reply cached. The document class probability, predicted class, the sub-class (pure AI or AI-paraphrased), and per-sentence flags were preserved. A text passes when GPTZero's predicted class is 'human'. Six human-written student essays from the ICNALE corpus were scored as controls. They are non-native writing, which is the population detectors over-flag, so their pass rate is a conservative bound on the detector's true-human rate.

3.5 Quality-gated pass rate and composite index

The quality-gated pass rate counts a GPTZero pass only when the same output also clears a usability gate, which is defined as the semantic cosine of at least 0.95, at least 90 percent of source citations preserved where the source had any, zero fabricated numbers or entities, zero injected characters, and length within tolerance. The share of outputs clearing the gate is reported on its own as the 'usable-output rate'. The gate is the same for every tool.

The composite index is an equal-weight mean of min-max normalised tool means over five components: rewrite depth, semantic faithfulness, distance from the human end of the spectrum, cleanliness (grammar errors and injected characters), and content integrity (the mean of named-entity retention, number retention, citation preservation and one minus number corruption). A second variant adds the quality-gated pass rate as a sixth component. No weights were tuned.

3.6 Statistics

Every tool mean has a 95 percent bootstrap confidence interval (5,000 resamples, seed 11). Since the design is paired, a Friedman test across tools uses the texts every tool completed (complete-case n is 15, or 11 for citation metrics), and each tool is compared with TextPulse by a Wilcoxon signed-rank test on pairwise-complete texts, with raw p-values and Benjamini-Hochberg false-discovery q-values both reported.

4. Every humanizer moves text toward the human end

Tables 1 to 3 present the per-tool means. Figures F1 to F11 show the distributions.

Tool	n	Spectrum	Cosine	Rewrite	FK shift	AI vocab /1k
TextPulse	48	29	0.971	0.52	-3.6	-39
Grammarly	47	26	0.982	0.35	-0.9	-32
QuillBot	22	106	0.951	0.68	-2.4	-31
Humanize AI Pro	47	52	0.976	0.43	-1.8	-31
StealthWriter	34	25	0.982	0.42	-1.6	-34
Walter Writes	48	51	0.959	0.60	-1.5	-10
HIX Bypass	19	85	0.981	0.39	-0.2	-19
GPTinf	48	53	0.960	0.52	-2.1	-30
UnAIMyText	48	29	0.964	0.54	-1.3	-38
Phrasly	47	57	0.962	0.54	-3.2	-38
Clever Humanizer	24	52	0.977	0.41	-1.5	-26
Untreated texts	48	138	1.000	0.00	0.0	0

Table 1. Style and faithfulness (per-tool means). Spectrum: position on the stylometric axis, human anchor median 0, AI anchor median 100. Cosine: BGE document similarity to the source. Rewrite: amount of words changed compared to original text. FK shift: Flesch-Kincaid grade change, a decrease is the desirable direction. AI vocab: change in AI-leaning lexicon words per 1,000 words, output minus source. Confidence intervals are in the publicly released results file.

Tool	n	Entities	Numbers	Citations	Fabricated	Grammar +/1k
TextPulse	48	0.95	0.87	0.99	0.48	0.65
Grammarly	47	0.99	0.85	1.00	0.06	4.91
QuillBot	22	0.61	0.51	0.29	0.14	1.60
Humanize AI Pro	47	0.89	0.90	0.98	0.30	0.75

Tool	n	Entities	Numbers	Citations	Fabricated	Grammar +/1k
StealthWriter	34	0.93	0.88	0.96	0.88	1.85
Walter Writes	48	0.81	0.52	0.70	4.44	1.73
HIX Bypass	19	0.94	0.75	0.93	2.74	3.34
GPTinf	48	0.76	0.75	0.63	0.38	2.51
UnAIMyText	48	0.85	0.68	0.97	1.60	2.42
Phrasly	47	0.78	0.83	0.88	2.79	3.58
Clever Humanizer	24	0.97	0.92	0.98	0.50	1.58
Untreated texts	48	1.00	1.00	1.00	0.00	0.00

Table 2. Content integrity, per-tool means. Entities: share of the source's named entities found in the output. Numbers, Citations: share of verified ground truth preserved in the output. Fabricated: invented numbers and named entities per output. Grammar: LanguageTool API errors introduced per 1,000 words.

Tool	n	Usable	GPTZero pass	Gated pass
TextPulse	48	67%	18.8%	10.4%
Grammarly	47	85%	0.0%	0.0%
QuillBot	22	0%	22.7%	0.0%
Humanize AI Pro	47	77%	2.1%	0.0%
StealthWriter	34	38%	0.0%	0.0%
Walter Writes	48	4%	60.4%	0.0%
HIX Bypass	19	37%	0.0%	0.0%
GPTinf	48	33%	14.6%	2.1%
UnAIMyText	48	21%	12.5%	0.0%
Phrasly	47	17%	38.3%	2.1%
Clever Humanizer	24	67%	16.7%	0.0%
Untreated texts	48		0.0%	

Table 3. Usability and detection (per-tool rates). Usable: portion of outputs passing the quality gate (i.e., cosine >= 0.95, citations preserved, no fabrications, no hidden characters, length within tolerance). GPTZero pass: share of outputs classified 'human'. Gated pass: GPTZero pass counted only on usable outputs.

The 48 untreated (raw AI) texts have a mean position of 138 on the spectrum, exceeding the AI anchor median of 100. Every humanizer moved its outputs toward the human end (Figure F1).

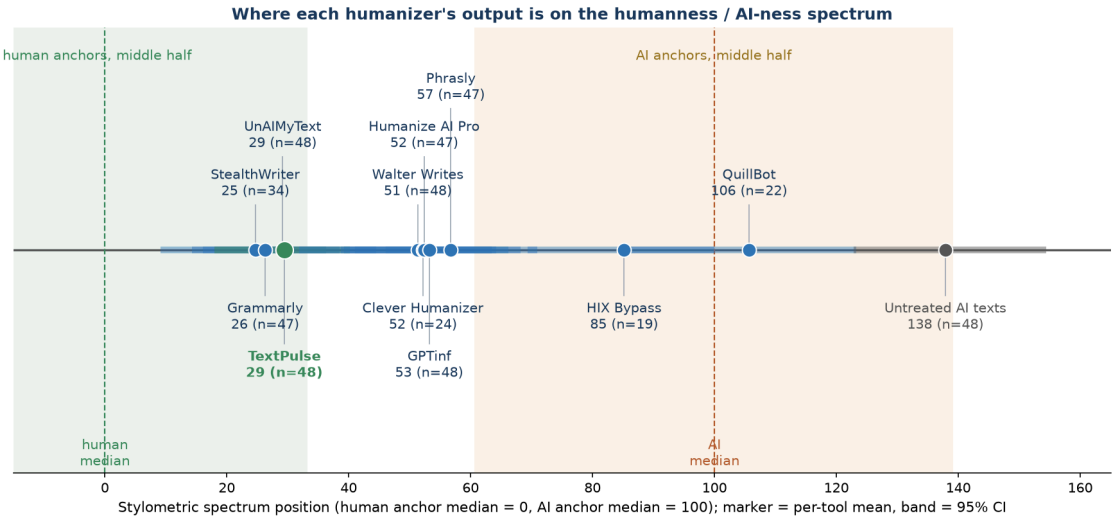

Figure F1. Every tool and the raw AI texts on the humanness / AI-ness spectrum, drawn as one scale. The marker is the per-tool mean, the band its 95 percent confidence interval. The shaded regions are the middle half of the human anchors and of the AI anchors.

Four tools reach the human range, defined as the middle half of the human anchors (-36 to 33 on the axis; Figure F1). TextPulse and UnAIMyText are tied at 29 (both n=48; TextPulse 95 percent CI 18 to 40), followed by Grammarly (26, n=47) and StealthWriter (25, n=34). The differences among the four are insignificant (raw p 0.50 to 0.64 against TextPulse), so the group is a tie on position, and within it TextPulse is the only tool that rewrote more than half of the source words to reach this region. Walter Writes (51), Clever Humanizer (52), Humanize AI Pro (52), GPTinf (53) and Phrasly (57) stop roughly halfway, outside the human range. HIX Bypass (85) and QuillBot (106) remain at the AI end. QuillBot's truncated outputs are more AI-like than the AI anchor median.

The features that move a text along this axis are concrete. Mean word length and the long-word rate are the two strongest ones. For example, a rewrite that turns "used" into "utilized" and "help" into "facilitate" increases both features. Lexical diversity increases with formal substitution for the same reason. Nominalizations ("the implementation of" for "implementing") and -ly adverbs ("significantly", "notably") load toward AI. Sentence-length burstiness loads toward human. For instance, a paragraph of sentences that all contain 22 to 26 words is flat and uniform, while a paragraph that mixes a 9-word sentence with a 31-word one is 'bursty'. Opener repetition loads toward human as well, since AI model output avoids starting consecutive sentences using the same openers, while humans prefer this.

5. Deep rewriting vs. faithfulness

Semantic faithfulness is generally high for most tools. Grammarly and StealthWriter (0.982), HIX Bypass (0.981), Clever Humanizer (0.977), Humanize AI Pro (0.976) and TextPulse (0.971, CI 0.965 to 0.975) all keep the document meaning close to the original source text. UnAIMyText (0.964), Phrasly (0.962), GPTinf (0.960) and Walter Writes (0.959), and QuillBot (0.951) are lower.

Grammarly changed 35 percent of source words and kept 36 percent of source trigrams verbatim. HIX Bypass changed 39 percent, Clever Humanizer 41 percent, StealthWriter 42 percent, and Humanize AI Pro 43 percent. TextPulse changed 52 percent and retained only 27 percent of trigrams, GPTinf 52 percent, Phrasly 54 percent, UnAIMyText 54 percent, and Walter Writes 60 percent. QuillBot (68 percent) changed the most and also lost the most content. The Friedman test on rewrite depth is significant (chi-square 95.2, $p < 0.001$, n=15), and TextPulse rewrites significantly more than Grammarly, StealthWriter, Humanize AI Pro, HIX Bypass and Clever Humanizer (q of 0.001 or below in each case).

These two measures together define the tool's position. Among tools that rewrote at least half the text, TextPulse has the highest faithfulness (0.971 against 0.964 or lower). Among tools with faithfulness of 0.97 or higher, it has the deepest rewrite (52 percent against 43 percent or lower). No other tool is in both groups.

6. Content integrity and fabrication rate

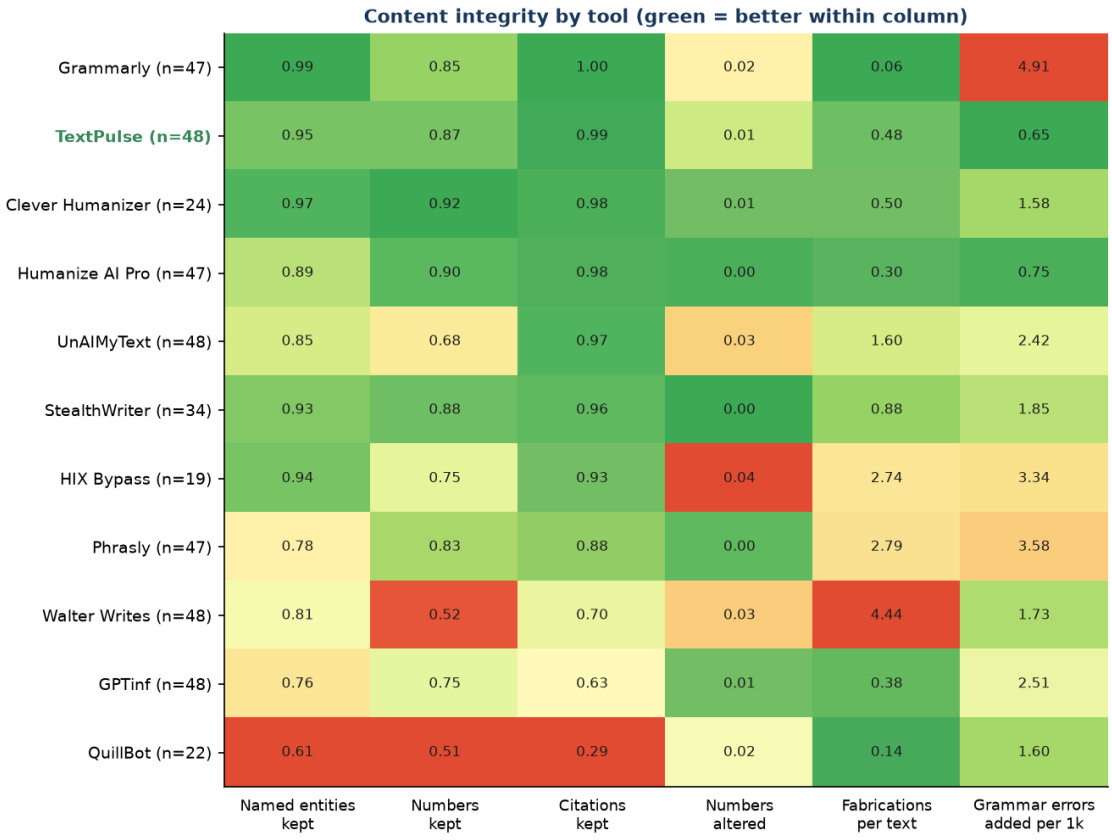

Figure F6. Content integrity by tool. Green is better within each column.

Named-entity retention is high for most tools, namely, Grammarly (0.99), Clever Humanizer (0.97), TextPulse (0.95, CI 0.87 to 1.00), HIX Bypass (0.94), StealthWriter (0.93), Humanize AI Pro (0.89), UnAIMyText (0.85), Walter Writes (0.81), Phrasly (0.78), GPTinf (0.76) and QuillBot (0.61). The differences among TextPulse and the light rewriters are not significant. TextPulse keeps more entities than Walter Writes, Phrasly, GPTinf and QuillBot at raw p from 0.007 to 0.024 (q 0.055 to 0.059); the Friedman test across tools is at p = 0.058 on the 12 complete-case texts that contain an entity. The entities lost are mostly units dropped when a statistic is reworded (“Cohen’s d” written as an effect size, “Celsius” written as degrees), author names dropped with a citation, and species, epoch or instrument names replaced by a description.

Number retention is highest for Clever Humanizer (0.92), Humanize AI Pro (0.90), StealthWriter (0.88) and TextPulse (0.87, CI 0.85 to 0.90), and lowest for Walter Writes (0.52) and QuillBot (0.51). Number corruption, the lower-bound rate of values changed in surviving context, is below 4 percent for every tool.

Citation preservation is near complete for Grammarly (1.00), TextPulse (0.994, CI 0.981 to 1.00), Clever Humanizer (0.98), Humanize AI Pro (0.98), UnAIMyText (0.97) and StealthWriter (0.96), and drops to 0.93 for HIX Bypass, 0.88 for Phrasly, 0.70 for Walter Writes, 0.63 for GPTinf and 0.29 for QuillBot. TextPulse converts parenthetical citations to narrative ones and merges adjacent numeric citations, and the matching rules count those as preserved.

In terms of fabrication, Walter Writes adds 4.4 numbers or named entities per output that do not exist in the source (4.0 numbers, 0.44 entities); Phrasly adds 2.8 (0.58 entities, the highest), HIX Bypass 2.7, UnAIMyText 1.6 and StealthWriter 0.9. Clever Humanizer (0.50), TextPulse (0.48, of which 0.02 entities), GPTinf (0.38), Humanize AI Pro (0.30), QuillBot (0.14) and Grammarly (0.06) add the least. Fabricated items include invented sample sizes, percentages and confidence bounds, and invented author names and institutions. Walter Writes fabricates significantly more than TextPulse (median difference 2.5 items per text, q below 0.001), as do Phrasly, HIX Bypass and UnAIMyText (q from 0.02 to below 0.001). TextPulse’s fabricated entity rate of one in fifty outputs is the lowest of any tool that rewrites more than 40 percent of the text.

7. Register, grammar and length

Every tool lowered the Flesch-Kincaid grade. TextPulse lowered it by 3.6 (CI 3.2 to 4.1), Phrasly by 3.2, QuillBot by 2.4, GPTinf by 2.1, Humanize AI Pro by 1.8, StealthWriter by 1.5, Walter Writes by 1.5, Clever Humanizer by 1.5, UnAIMyText by 1.3, Grammarly by 0.9, and HIX Bypass by 0.2. AI Model-generated academic text reads at a higher grade than published human academic prose, so a decrease toward the human range is the intended direction, and TextPulse’s decrease is the largest of any tool.

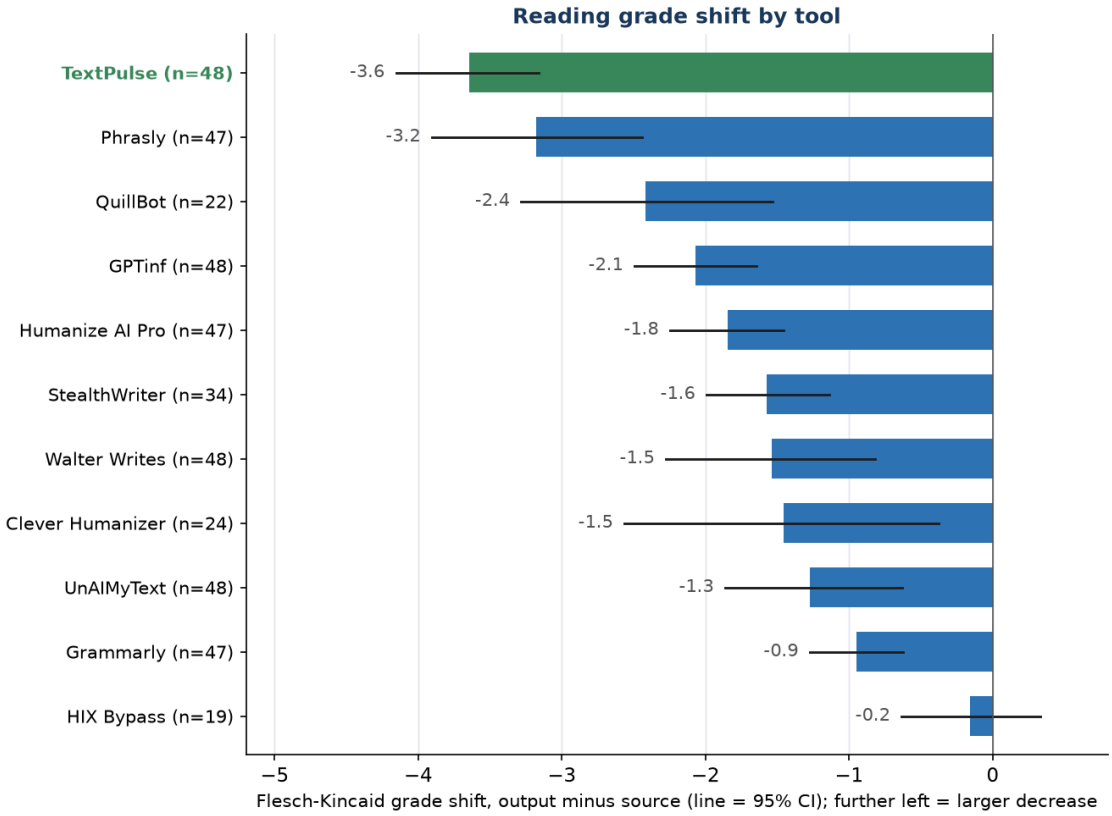

Figure F11. Flesch-Kincaid grade shift per tool, output minus source, with 95 percent confidence intervals. A larger decrease moves the text further toward the simplified formal register of human academic writing.

Grammar errors introduced per 1,000 words are lowest for TextPulse (0.65, CI 0.22 to 1.20) and Humanize AI Pro (0.75), and highest for Grammarly (4.9), Phrasly (3.6) and HIX Bypass (3.3). Grammarly's rate is significantly higher than TextPulse's (q below 0.001). TextPulse introduces the fewest grammar errors of any tool compared.

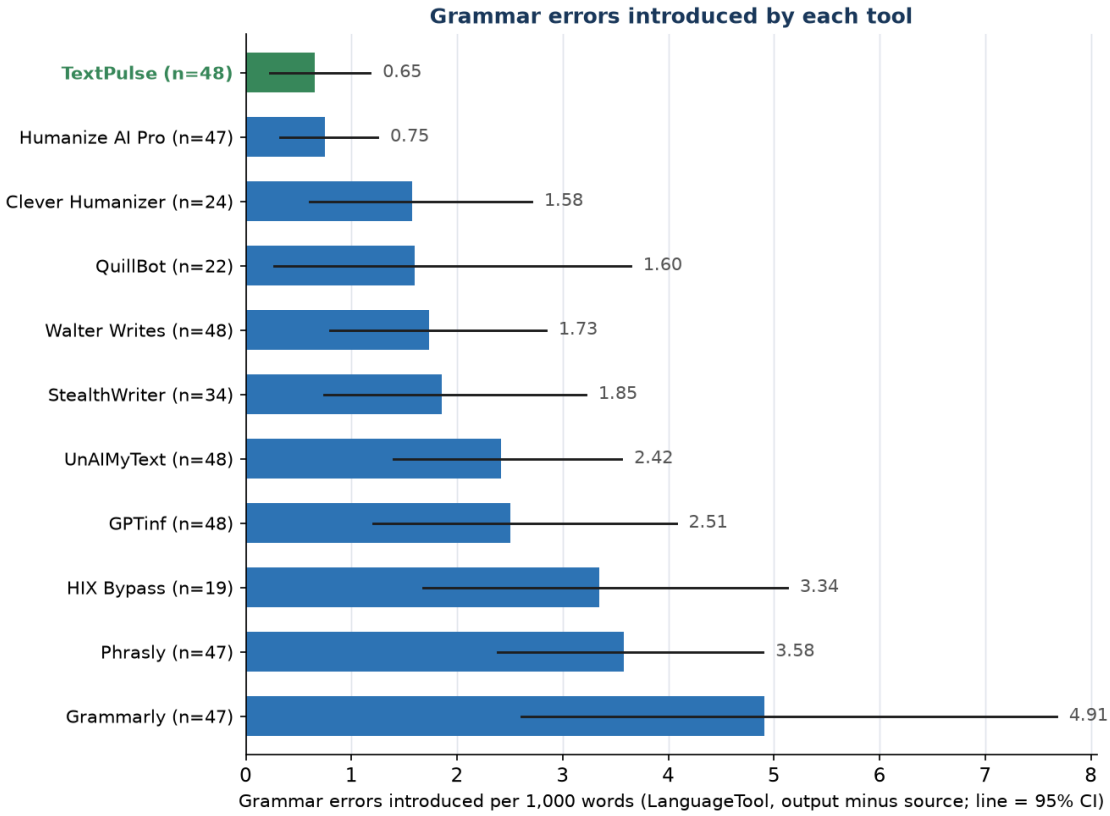

Figure F10. Grammar errors introduced per 1,000 words by each tool, LanguageTool count in the output minus the source, with 95 percent confidence intervals.

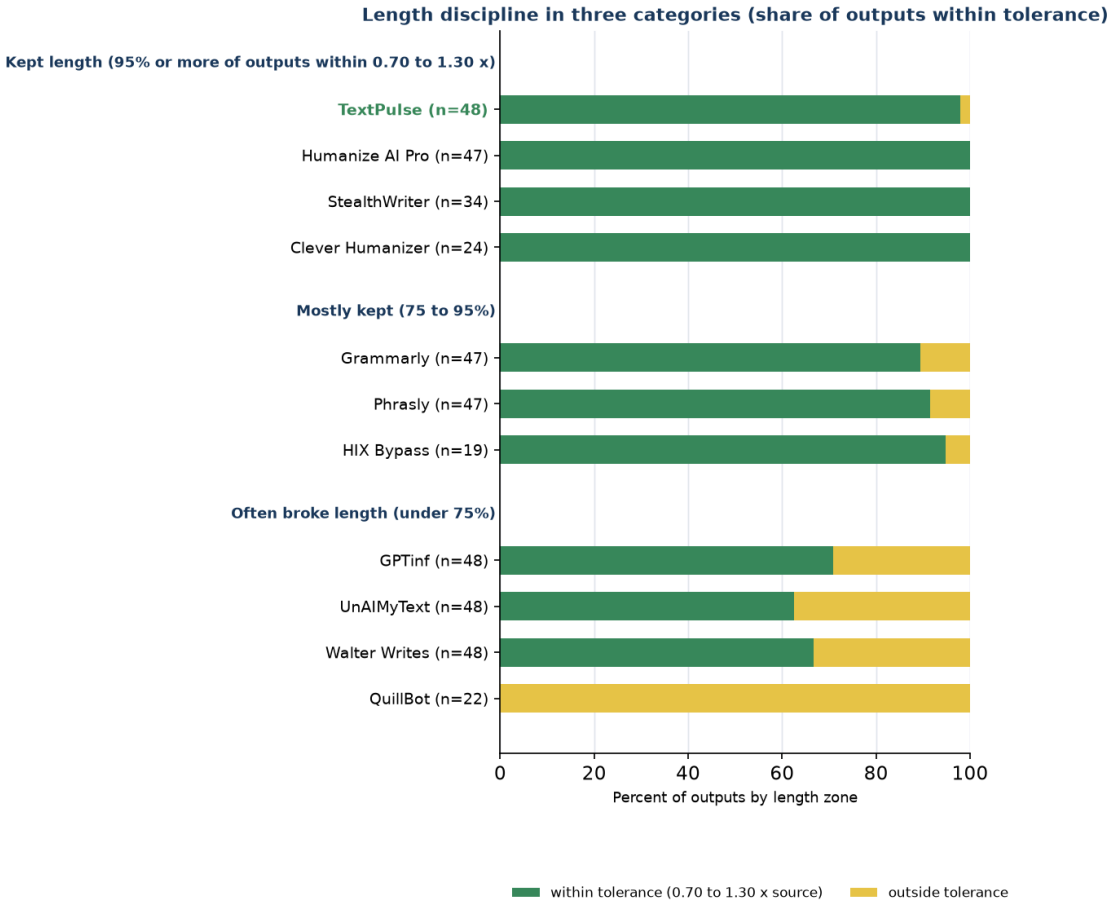

Figure F5. Length discipline in three categories. Bars give the share of outputs in each length zone: green within tolerance (0.70 to 1.30 times the source length), yellow outside it. Within a category tools are listed by sample size.

Length discipline falls into three categories (Figure F5). Four tools kept length, with 95 percent or more of outputs within tolerance: TextPulse (47 of 48, mean ratio 1.14), Humanize AI Pro (47 of 47, 1.06), StealthWriter (34 of 34, 1.10) and Clever Humanizer (24 of 24, 1.06). Three tools mostly kept length: Grammarly (42 of 47, ratio 1.25, most outputs expanded toward the upper limit), Phrasly (43 of 47) and HIX Bypass (18 of 19). Four tools often broke length: GPTinf (34 of 48, 0.81, in part because of a word cap), UnAIMyText (30 of 48, 1.25), Walter Writes (32 of 48, 1.33) and QuillBot (0 of 22, all outside tolerance at a mean ratio of 0.43).

None of the models under study injected any hidden characters into the output text.

8. Every tool removes AI vocabulary

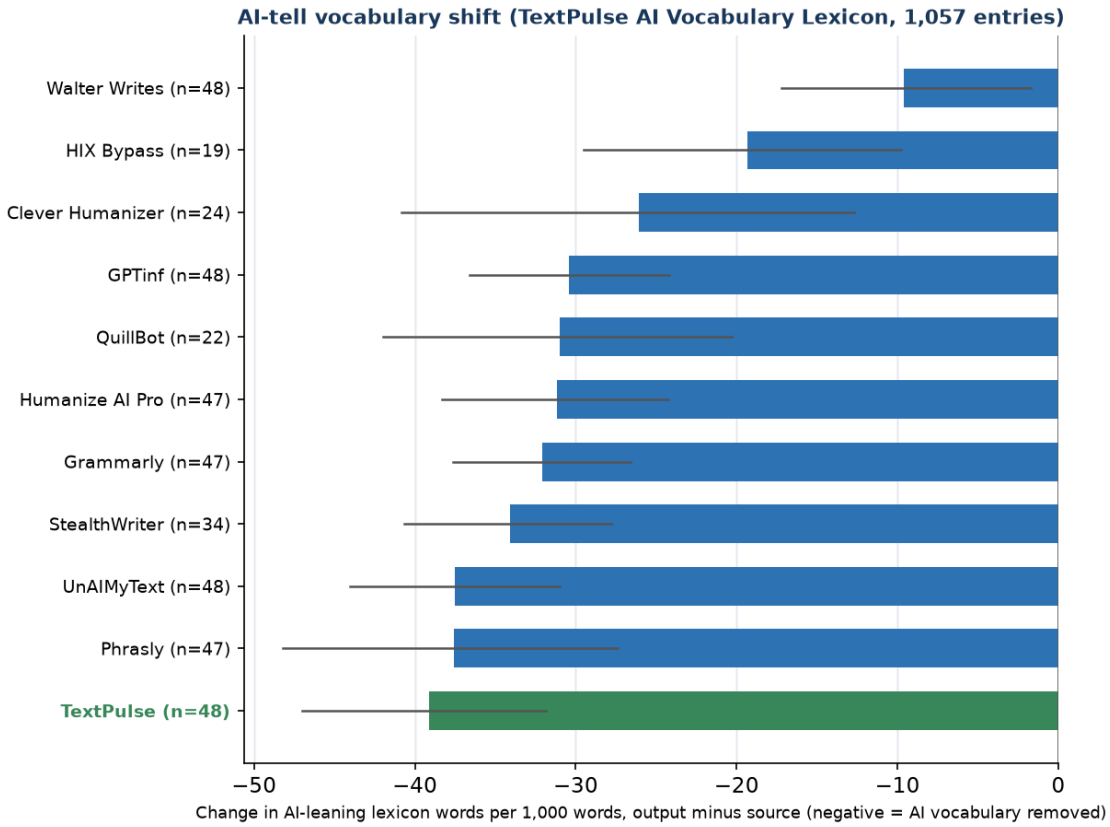

Figure F7. Change in AI-leaning vocabulary per 1,000 words, output minus source.

Every tool reduced AI-typical vocabulary (Figure F7). The change in AI-leaning lexicon words per 1,000 words is -39 for TextPulse (CI -47 to -32), -38 for Phrasly and UnAIMyText, -34 for StealthWriter, -32 for Grammarly, -31 for Humanize AI Pro and QuillBot, -30 for GPTinf, -26 for Clever Humanizer, -19 for HIX Bypass and -10 for Walter Writes. The vocabulary index, which also credits human-leaning words added, gives the same order with TextPulse at -80 and Walter Writes at -57. TextPulse removes significantly more AI-leaning vocabulary than Walter Writes, HIX Bypass, Clever Humanizer, Grammarly, GPTinf and Humanize AI Pro (q from 0.014 to below 0.001), and about the same as Phrasly, UnAIMyText, StealthWriter and QuillBot. It removes the most of any tool compared.

9. Detector outcomes

9.1 Turnitin

Turnitin returns one AI percentage per document and prints an asterisk under 20 percent. With one document per tool, every document scored above the threshold (Table 4, Figure F9). TextPulse and StealthWriter score lowest at 24 percent, then Walter Writes (28), HIX Bypass (35), Humanize AI Pro (39), Phrasly (46), UnAIMyText (57), Clever Humanizer (61), Grammarly and GPTinf (68), and QuillBot (74).

Tool	n	Turnitin AI %
TextPulse	48	24%
StealthWriter	34	24%
Walter Writes	48	28%
HIX Bypass	19	35%
Humanize AI Pro	47	39%
Phrasly	47	46%
UnAIMyText	48	57%
Clever Humanizer	24	61%
GPTinf	48	68%
Grammarly	47	68%
QuillBot	22	74%

Table 4. Turnitin document-level AI percentage per tool. Each document holds only that tool's outputs; lower is less flagged.

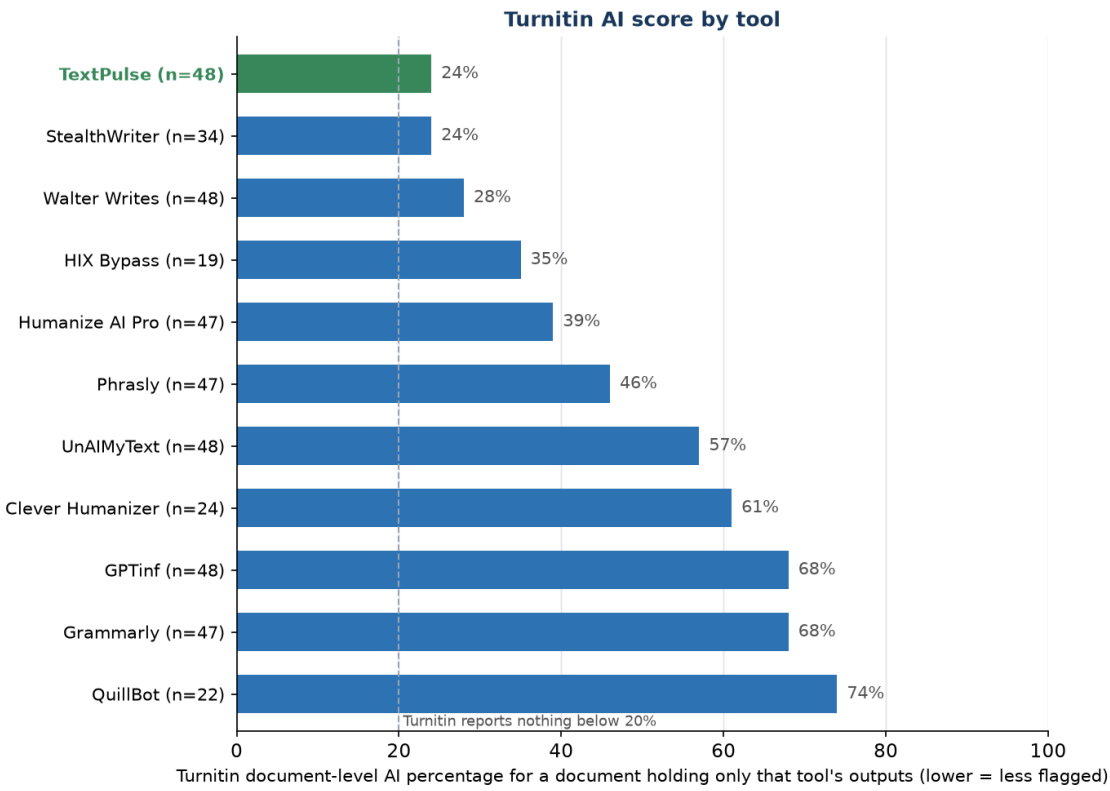

Figure F9. Turnitin document-level AI percentage by tool. Each bar is one document holding only that tool's outputs; lower is less flagged.

Turnitin's score is a document-level figure, so it gives one number per tool and no per-text distribution or confidence interval. Read as a ranking, it agrees with GPTZero at the two ends: the tools that rewrite deeply while keeping the text intact are flagged least, and QuillBot's truncated outputs are flagged most. It disagrees in the middle, where Grammarly, which passes nothing on GPTZero, is flagged at 68 percent, and Walter Writes, which passes GPTZero most often, is flagged at 28 percent. TextPulse has the lowest Turnitin score of any tool, tied with StealthWriter, while rewriting 52 percent of the words against StealthWriter's 42.

9.2 GPTZero

GPTZero classified all 48 untreated texts as AI with probability 1.000 and every one as pure AI rather than paraphrased. Five of the six human control essays were classified human (one was flagged at probability 1.000). Across tools, raw pass rates were: Walter Writes 60.4 percent (29 of 48), Phrasly 38.3, QuillBot 22.7, TextPulse 18.8 (9 of 48, CI 8.3 to 29.2), Clever Humanizer 16.7, GPTinf 14.6, UnAIMyText 12.5, Humanize AI Pro 2.1, and zero for Grammarly, StealthWriter and HIX Bypass. The Friedman test is significant (chi-square 44.7, p below 0.001).

10. Passing a detector without breaking the text

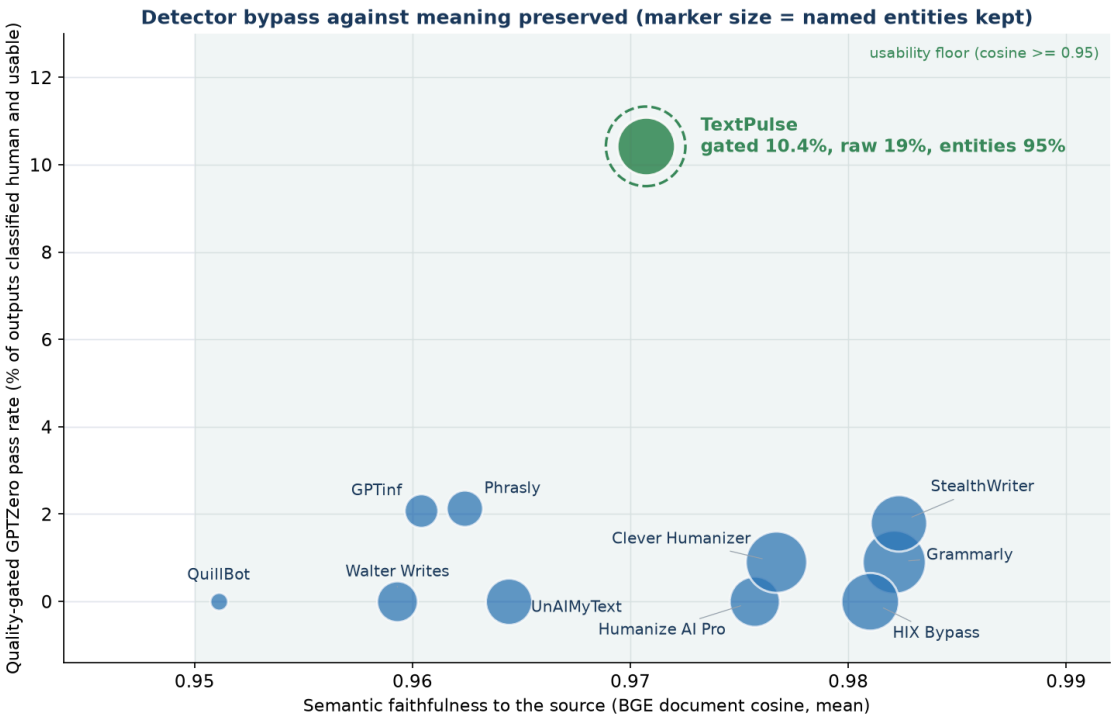

Figure F2. Quality-gated GPTZero pass rate against semantic faithfulness. Marker area scales with the share of named entities kept, so a tool that keeps the text intact is drawn large; markers that share a position are stacked slightly. Table 5 gives the values.

Tool	n	Gated pass	Raw pass	Cosine	Entities kept
TextPulse	48	10.4%	19%	0.971	95%
Phrasly	47	2.1%	38%	0.962	78%
GPTinf	48	2.1%	15%	0.960	76%
Walter Writes	48	0.0%	60%	0.959	81%
QuillBot	22	0.0%	23%	0.951	61%
Clever Humanizer	24	0.0%	17%	0.977	97%
UnAIMyText	48	0.0%	12%	0.964	86%
Humanize AI Pro	47	0.0%	2%	0.976	89%
StealthWriter	34	0.0%	0%	0.982	93%
Grammarly	47	0.0%	0%	0.982	99%
HIX Bypass	19	0.0%	0%	0.981	94%

Table 5. Values plotted in Figure F2, in order of quality-gated pass rate.

The three tools with the highest raw pass rates are three of the four tools with the lowest ‘usable’ output rates. Walter Writes passes 60 percent of its outputs and only 4 percent of them clear the usability gate; Phrasly passes 38 percent, with 17 percent usable; QuillBot passes 23 percent with none usable. Figure F4 shows the reason. Walter Writes fails on fabrication in 90 percent of outputs and on citations in 38 percent, Phrasly on fabrication in 74 percent, QuillBot on length in 100 percent and citations in 64 percent. The tools that keep the text intact most often, Grammarly (85 percent usable), Clever Humanizer (79 percent), Humanize AI Pro (77 percent) and TextPulse (67 percent), pass GPTZero at 0, 17, 2 and 19 percent.

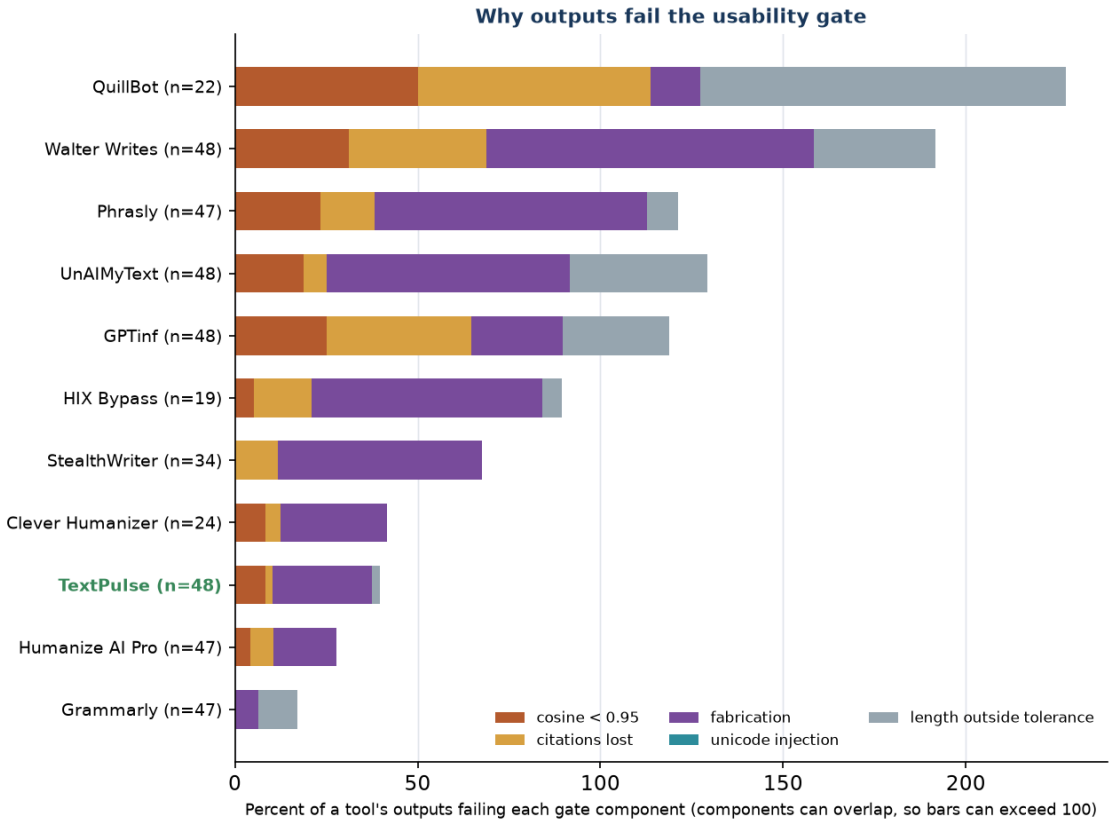

Figure F4. Why outputs fail the usability gate, by component. Components overlap, so bars can exceed 100.

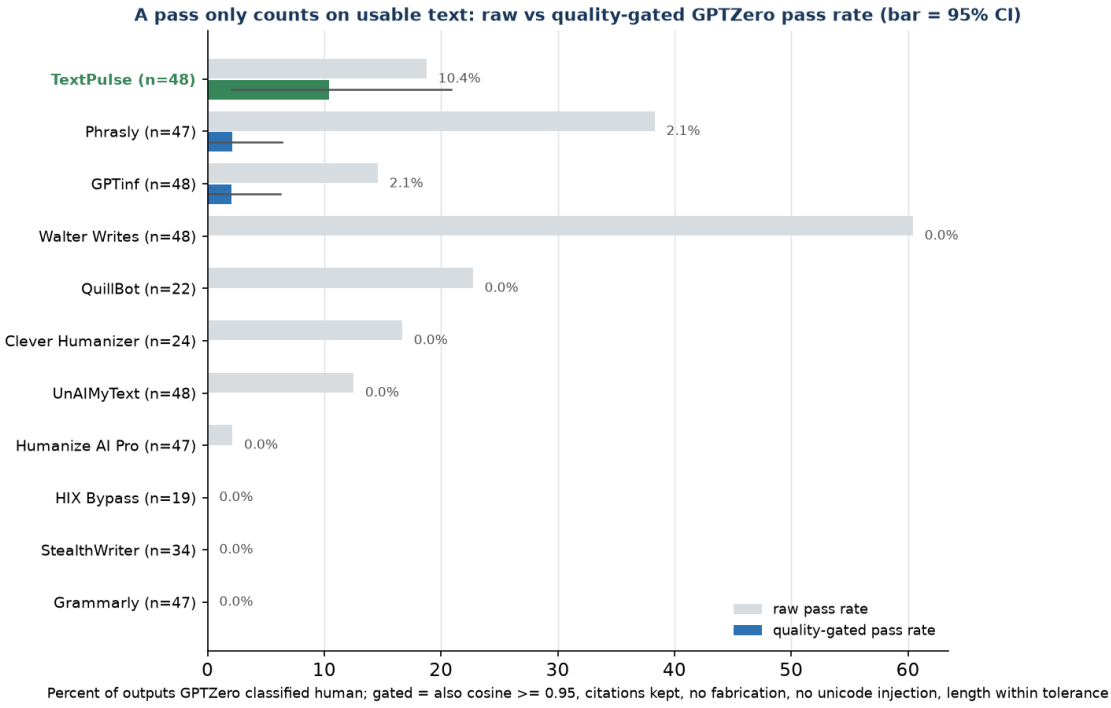

Figure F3. Raw and quality-gated GPTZero pass rate, with 95 percent confidence intervals on the gated rate.

Counting only passes on usable outputs (Figure F3), TextPulse has the highest gated rate at 10.4 percent (5 of 48, CI 2.1 to 18.8), then Phrasly and GPTinf at 2.1 percent (1 each), and zero for the other eight tools including Walter Writes. TextPulse's gated rate is higher than Grammarly's, Humanize AI Pro's, Walter Writes' and UnAIMyText's at raw p = 0.025 and StealthWriter's at p = 0.046.

Figure F2 plots the gated pass rate against faithfulness. TextPulse is the highest point on the chart. Walter Writes and Phrasly, the two tools with the highest raw pass rates, are at or near zero on the gated rate because almost all of their raw passes came from outputs that failed the gate. The light rewriters are at the far right with no passes.

11. Composite index

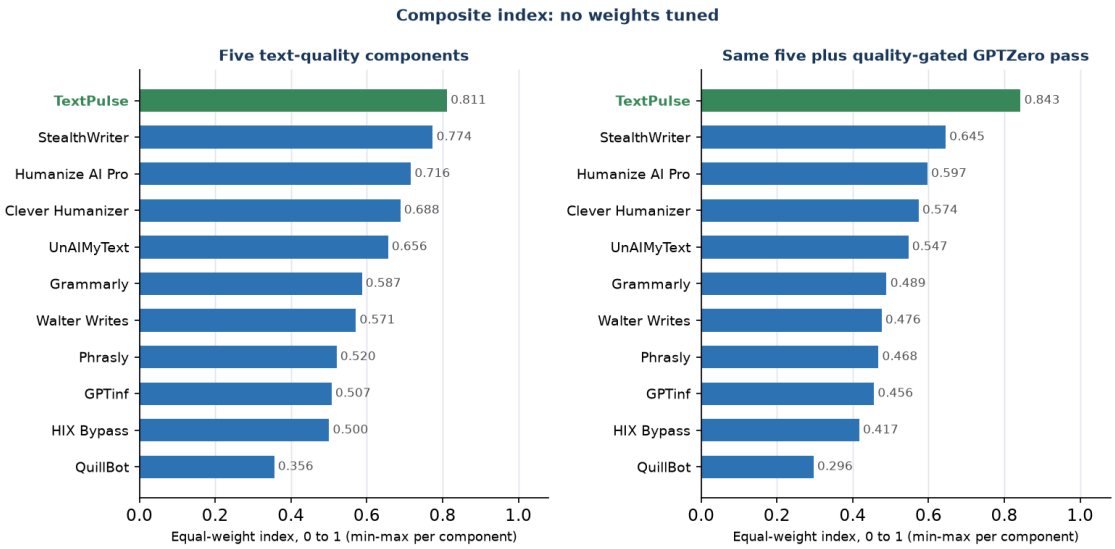

Figure F8. Composite index, five text-quality components and the same five plus the quality-gated detector pass. No weights were tuned.

On the five text-quality components, TextPulse scores 0.811, StealthWriter 0.774, Humanize AI Pro 0.716, Clever Humanizer 0.688, UnAIMyText 0.656, Grammarly 0.587, Walter Writes 0.571, Phrasly 0.520, GPTinf 0.507, HIX Bypass 0.500 and QuillBot 0.356. Adding the gated pass rate as a sixth component gives TextPulse 0.843 and StealthWriter 0.645, with the rest of the order nearly unchanged (Figure F8). TextPulse is first because it is the only tool without a weak component: it rewrites deeply, keeps meaning, is near the human end, adds few errors, and preserves citations and numbers. StealthWriter and Grammarly match or beat it on faithfulness and content, and lose on rewrite depth and, for Grammarly, on grammar. Walter Writes and Phrasly beat it on raw detector pass and lose on fabrication and citations. Readers who weight the components differently can recompute the index from the released data.

12. Discussion

The central result is a trade-off. Across eleven tools, the outputs most likely to pass a detector are the outputs that have changed the underlying meaning of the text. Walter Writes, the tool with the highest GPTZero pass rate, adds more than four invented numbers or names per 300-word text and loses 30 percent of the citations. Phrasly, second on pass rate, adds nearly three invented items and 3.6 grammar errors per 1,000 words. What use is a high pass rate if the text contains errors?

TextPulse was designed to rewrite in a human style, and prioritizes preserving grammar, formality, entities, citations and meaning, while passing detection. It moves text to the human side of the spectrum with the top group, rewrites more than half the words, keeps 99 percent of citations and 87 percent of numbers, adds only 0.65 grammar errors per 1,000 words, fabricates almost no entities, and removes more AI-leaning vocabulary than any tool that keeps the text intact.

On detection it is first among tools that keep the text intact, at a gated rate of 10 percent with a confidence interval from 2 to 19 percent, and it is under Turnitin's reporting threshold. The other tools prioritise passing detection at the expense of those factors and show a quick degradation in grammar, formality, entities and meaning, as the figures above demonstrate.

The detector results have two further implications. Turnitin's document scores place TextPulse lowest, tied with StealthWriter at 24 percent, and place the light rewriter Grammarly and the deep rewriter GPTinf together at 68 percent, so a document-level detector score on its own does not track rewrite depth or content preservation. GPTZero flags most humanized academic text, labels it as paraphrased AI, and flagged one of six human control essays. Any humanizer pass rate should be read against that false-positive rate, and the same sentence-level flags that catch humanized text will also catch some human writing.

The spectrum position and the vocabulary shift show that the tools change style in a measurable, consistent direction. Every tool reduced AI-leaning vocabulary, and the reduction is largest for the tools that rewrite most. Every tool lowered the reading grade. These changes are what a stylometric detector responds to, and they are also the changes that move a text toward the register of published human academic writing. The two goals coincide when a tool keeps the content.

13. Limitations and conclusion

The texts were written by four AI generators to a fixed prompt, and the ground truth is what those generators reported and we verified. Real drafts are longer and less uniform. Each tool was run once, at one setting, on the dates recorded. Vendors update tools without notice, and these results must be considered a snapshot. Per-tool n is unequal because free tiers limit throughput, so the complete-case Friedman tests rest on 15 texts, and several tools have intervals wide enough to overlap most of the field. Turnitin gives one document-level score per tool with no per-text distribution, and GPTZero is one detector with one model version.

Eleven AI text humanizers rewrote the same 48 AI-generated academic texts. All of them moved the text toward the human end of a frozen stylometric spectrum, reduced AI-leaning vocabulary

and lowered the reading grade. They differed sharply in what they preserved. Tools that pass GPTZero most often fabricate numbers and names, drop citations and add grammar errors, while tools that keep the text intact pass less often. TextPulse is the exception on both counts. It is first on the quality-gated pass rate, first on the composite index with or without the detector, and in the top group on every content measure while rewriting more than half of the text. The measure that matters for academic writing is the combination, and by that measure the field has one tool that keeps the text intact while passing at a measurable rate, and a market that mostly does either one or the other.

Data availability

The 48 source texts with verified ground truth, all 432 humanized outputs from the eleven tools, the GPTZero responses for all 480 texts and six controls, the Turnitin document scores, per-text scores, analysis results, figures, and every script are released at textpulse.ai/research and archived on Zenodo (DOI 10.5281/zenodo.22170060).

References

Ardito, C. G. (2025). Generative AI detection in higher education assessments. New Directions for Teaching and Learning. https://doi.org/10.1002/tl.20624

Artemova, E., Lucas, J., Venkatraman, S., Lee, J., Tilga, S., Uchendu, A., & Mikhailov, V. (2025). Beemo: Benchmark of expert-edited machine-generated outputs. In Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2025). https://aclanthology.org/2025.naacl-long.357/

Astarita, S., Kruk, S., Reerink, J., & Gómez, P. (2024). Delving into the utilisation of ChatGPT in scientific publications in astronomy. arXiv. https://arxiv.org/abs/2406.17324

Baidya, M. S., Baidya, S. S., & Chawla, C. (2026). Detecting the machine: A comprehensive benchmark of AI-generated text detectors across architectures, domains, and adversarial conditions. arXiv. https://arxiv.org/abs/2603.17522

Bailey, J. (2025, August 27). Turnitin launches anti-AI humanizer feature. Plagiarism Today. https://www.plagiarismtoday.com/2025/08/27/turnitin-launches-anti-ai-humanizer-feature/

Bao, G., Rong, L., Zhao, Y., Zhou, Q., & Zhang, Y. (2025). Decoupling content and expression: Two-dimensional detection of AI-generated text. arXiv. https://arxiv.org/abs/2503.00258

Bao, G., Zhao, Y., Teng, Z., Yang, L., & Zhang, Y. (2024). Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024). https://arxiv.org/abs/2310.05130

Barrot, J. S., & Aranda, M. R. R. (2025). Efficacy of AI-text detection tools in distinguishing student-produced, AI-edited, and AI-generated essays. Technology, Knowledge and Learning. https://doi.org/10.1007/s10758-025-09884-0

Bassett, M. A., Bradshaw, W., Bornsztejn, H., Hogg, A., Murdoch, K., Pearce, B., & Webber, C. (2026). Heads we win, tails you lose: AI detectors in education. Journal of Higher Education Policy and Management. https://doi.org/10.1080/1360080X.2026.2622146

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. Journal of Applied Learning and Teaching, 6(2), 94-104. https://doi.org/10.37074/jalt.2023.6.2.12

Chaka, C. (2024a). Accuracy pecking order: How 30 AI detectors stack up in detecting generative artificial intelligence content in university English L1 and English L2 student essays. Journal of Applied Learning and Teaching, 7(1), 127-139. https://doi.org/10.37074/jalt.2024.7.1.33

Chaka, C. (2024b). Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review. Journal of Applied Learning and Teaching, 7(1), 115-126. https://doi.org/10.37074/jalt.2024.7.1.14

Cheng, Y., Guo, Z., Saha, A., Kadur, S., Sedhain, G., & Feizi, S. (2025). Adversarial paraphrasing: A universal attack for humanizing AI-generated text. In Advances in Neural Information Processing Systems 38 (NeurIPS 2025). https://arxiv.org/abs/2506.07001

Crothers, E., Japkowicz, N., & Viktor, H. L. (2023). Machine-generated text: A comprehensive survey of threat models and detection methods. IEEE Access, 11, 70977-71002. https://doi.org/10.1109/ACCESS.2023.3294090

David, I., & Gervais, A. (2025). AuthorMist: Evading AI text detectors with reinforcement learning. arXiv. https://arxiv.org/abs/2503.08716

Desaire, H., Chua, A. E., Isom, M., Jarosova, R., & Hua, D. (2023). Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. Cell Reports Physical Science, 4(6), 101426. https://doi.org/10.1016/j.xcrp.2023.101426

Doughman, J., Afzal, O. M., Toyin, H. O., Shehata, S., Nakov, P., & Talat, Z. (2025). Exploring the limitations of detecting machine-generated text. In Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025) (pp. 4274-4281). https://aclanthology.org/2025.coling-main.288/

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

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. https://doi.org/10.1007/s40979-023-00140-5

Emi, B., & Spero, M. (2024). Technical report on the Pangram AI-generated text classifier. arXiv. https://arxiv.org/abs/2402.14873

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

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. npj Digital Medicine, 6, 75. https://doi.org/10.1038/s41746-023-00819-6

Garland, N. A. (2026). AI detectors fail diverse student populations: A mathematical framing of structural detection limits. arXiv. https://arxiv.org/abs/2603.20254

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical detection and visualization of generated text. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 111-116). https://aclanthology.org/P19-3019/

Geng, M., & Trotta, R. (2024). Is ChatGPT transforming academics' writing style? arXiv. https://arxiv.org/abs/2404.08627

Glickenhaus, B., Thai, K., Russell, J., et al. (2026). Pangram 4 technical report. arXiv. https://arxiv.org/abs/2607.27183

Gu, Y., Li, S., & Hu, X. (2026). MASH: Evading black-box AI-generated text detectors via style humanization. In Findings of the Association for Computational Linguistics: ACL 2026. https://arxiv.org/abs/2601.08564

Hadra, M., Cambridge, K., & Mesbah, M. (2026). Evaluating the accuracy and reliability of AI content detectors in academic contexts. International Journal for Educational Integrity. https://doi.org/10.1007/s40979-026-00213-1

Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J., & Goldstein, T. (2024). Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text. In Proceedings of the 41st International Conference on Machine Learning (PMLR 235). https://proceedings.mlr.press/v235/hans24a.html

He, X., Shen, X., Chen, Z., Backes, M., & Zhang, Y. (2024). MGTBench: Benchmarking machine-generated text detection. In Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS 2024). https://doi.org/10.1145/3658644.3670344

Hu, B. (2026). Why artificial intelligence detectors could penalize academic writing. Nature Human Behaviour, 10, 439. https://doi.org/10.1038/s41562-026-02420-9

Hu, X., Chen, P.-Y., & Ho, T.-Y. (2023). RADAR: Robust AI-text detection via adversarial learning. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023).

https://arxiv.org/abs/2307.03838

Ibrahim, H., Liu, F., Asim, R., Battu, B., Benabderrahmane, S., Alhafni, B., et al. (2023). Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. Scientific Reports, 13, 12187. https://doi.org/10.1038/s41598-023-38964-3

Ibrahim, K. H., Al Otaibi, D., & Sibai, F. N. (2025). The robustness of AI-classifiers in the face of AI-assisted plagiarism: The case of Turnitin AI Content Detector. International Journal of Computer-Assisted Language Learning and Teaching, 15(1), 1-27. https://doi.org/10.4018/IJCALLT.372428

Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). Automatic detection of generated text is easiest when humans are fooled. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1808-1822). https://aclanthology.org/2020.acl-main.164/

Jabarian, B., & Imas, A. (2025). Artificial writing and automated detection (NBER Working Paper No. 34223). National Bureau of Economic Research. https://www.nber.org/papers/w34223

Jawahar, G., Abdul-Mageed, M., & Lakshmanan, L. V. S. (2020). Automatic detection of machine generated text: A critical survey. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 2296-2309). https://aclanthology.org/2020.coling-main.208/

Jovanović, N., Staab, R., & Vechev, M. (2024). Watermark stealing in large language models. In Proceedings of the 41st International Conference on Machine Learning (PMLR 235). https://proceedings.mlr.press/v235/jovanovic24a.html

Juzek, T. S., & Ward, Z. B. (2025). Why does ChatGPT “delve” so much? Exploring the sources of lexical overrepresentation in large language models. In Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025). https://aclanthology.org/2025.coling-main.426/

Kar, S. K., Bansal, T., Modi, S., & Singh, A. (2025). How sensitive are the free AI-detector tools in detecting AI-generated texts? A comparison of popular AI-detector tools. Indian Journal of Psychological Medicine, 47(3), 275-278. https://doi.org/10.1177/02537176241247934

Karr, J. A., Jr., Khvatskii, G., Hua, T., & Chawla, N. V. (2026). Why AI detection fails for academic integrity. arXiv. https://arxiv.org/abs/2608.11256

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. In Proceedings of the 40th International Conference on Machine Learning (PMLR 202, pp. 17061-17084). https://proceedings.mlr.press/v202/kirchenbauer23a.html

Kobak, D., González-Márquez, R., Horvát, E.-Á., & Lause, J. (2025). Delving into LLM-assisted writing in biomedical publications through excess vocabulary. Science Advances, 11(27), eadt3813. https://doi.org/10.1126/sciadv.adt3813

Koike, R., Kaneko, M., & Okazaki, N. (2024). OUTFOX: LLM-generated essay detection through in-context learning with adversarially generated examples. Proceedings of the AAAI Conference on Artificial Intelligence, 38(19), 21258-21266. https://doi.org/10.1609/aaai.v38i19.30120

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023). https://arxiv.org/abs/2303.13408

Li, Y., Li, Q., Cui, L., Bi, W., Wang, Z., Wang, L., Yang, L., Shi, S., & Zhang, Y. (2024). MAGE: Machine-generated text detection in the wild. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 36-53). https://aclanthology.org/2024.acl-long.3/

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., & Zou, J. Y. (2024). Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In Proceedings of the 41st International Conference on Machine Learning (PMLR 235). https://proceedings.mlr.press/v235/liang24b.html

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Liang, W., Zhang, Y., Wu, Z., et al. (2025). Quantifying large language model usage in scientific papers. Nature Human Behaviour, 9, 2599-2609. https://doi.org/10.1038/s41562-025-02273-8

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., & Zou, J. Y. (2024). Mapping the increasing use of LLMs in scientific papers. In Proceedings of the First Conference on Language Modeling (COLM 2024). https://arxiv.org/abs/2404.01268

Masrour, E., Emi, B., & Spero, M. (2025). DAMAGE: Detecting adversarially modified AI generated text. In Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect) at COLING 2025. https://aclanthology.org/2025.genaidetect-1.9/

Matsui, K. (2024). Delving into PubMed records: Some terms in medical writing have drastically changed after the arrival of ChatGPT. medRxiv. https://doi.org/10.1101/2024.05.14.24307373

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

Pedrotti, A., Papucci, M., Ciaccio, C., Miaschi, A., Puccetti, G., Dell'Orletta, F., & Esuli, A. (2025). Stress-testing machine generated text detection: Shifting language models writing style to fool detectors. In Findings of the Association for Computational Linguistics: ACL 2025. https://aclanthology.org/2025.findings-acl.156/

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2024). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. Journal of Academic Ethics, 22(1), 89-113. https://doi.org/10.1007/s10805-023-09492-6

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). Simple techniques to bypass GenAI text detectors: Implications for inclusive education. International Journal of Educational Technology in Higher Education, 21, 53. https://doi.org/10.1186/s41239-024-00487-w

Perrone, G., & Romano, S. P. (2026). ARB: A matched authorship-rewriting benchmark dataset for AI-text detector evaluation. arXiv. https://arxiv.org/abs/2607.29539

Ranganath, S., & Ramesh, A. (2026). StealthRL: Reinforcement learning paraphrase attacks for multi-detector evasion of AI-text detectors. arXiv. https://arxiv.org/abs/2602.08934

Russell, J., Karpinska, M., & Iyyer, M. (2025). People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 5342-5373). https://aclanthology.org/2025.acl-long.267/

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2025). Can AI-generated text be reliably detected? Stress testing AI text detectors under various attacks. Transactions on Machine Learning Research. https://openreview.net/forum?id=OOgsAZdFOt

Shi, Z., Wang, Y., Yin, F., Chen, X., Chang, K.-W., & Hsieh, C.-J. (2024). Red teaming language model detectors with language models. Transactions of the Association for Computational Linguistics, 12, 174-189. https://doi.org/10.1162/tacl_a_00639

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., & Wang, J. (2019). Release strategies and the social impacts of language models. arXiv. https://arxiv.org/abs/1908.09203

Sourati, Z., et al. (2026). The shrinking landscape of linguistic diversity in the age of large language models. Nature Human Behaviour. https://doi.org/10.1038/s41562-026-02550-0

Sun, Y., Liao, Y., & Ma, X. (2026). Trusting AI to detect AI? A systematic evaluation of the reliability and robustness of current AIGC detection tools for student academic work. Computers & Education. https://doi.org/10.1016/j.compedu.2026.105456

Tang, R., Chuang, Y.-N., & Hu, X. (2024). The science of detecting LLM-generated text. Communications of the ACM, 67(4), 50-59. https://doi.org/10.1145/3624725

TextPulse Research. (2026a). Do AI detectors agree? An inter-rater reliability study of commercial AI text detectors on academic writing. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22003419

TextPulse Research. (2026b). Do AI models invent references? A verification audit of citations in AI-generated academic text. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22010511

TextPulse Research. (2026c). What happens to citations when AI rewrites academic text? A large-scale paired audit. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22010530

TextPulse Research. (2026d). The vocabulary fingerprint of AI rewriting: Common words AI language models prioritize. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22028377

TextPulse Research. (2026e). Stylometric fingerprints of AI rewriting: Punctuation, syntax, and model attribution across 60,786 paired texts. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22040683

TextPulse Research. (2026f). Human versus AI text classification from stylometric features across 121,092 academic texts. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22048476

TextPulse Research. (2026g). Non-native English writing and the false positives of stylometric AI text detection. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22055658

Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Nikolenko, S., Burnaev, E., Barannikov, S., & Piontkovskaya, I. (2023). Intrinsic dimension estimation for robust detection of AI-generated texts. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023). https://arxiv.org/abs/2306.04723

Turnitin. (2025, August 27). Turnitin expands capabilities amid rising threats posed by AI bypassers [Press release]. https://www.turnitin.com/press/turnitin-expands-capabilities-amid-rising-threats-posed-by-ai-bypassers

Uchendu, A., Le, T., Shu, K., & Lee, D. (2020). Authorship attribution for neural text generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 8384-8395). https://aclanthology.org/2020.emnlp-main.673/

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. https://doi.org/10.1007/s40979-026-00226-w

Verma, V., Fleisig, E., Tomlin, N., & Klein, D. (2024). Ghostbuster: Detecting text ghostwritten by large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 1702-1717). https://aclanthology.org/2024.naacl-long.95/

Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. Open Information Science, 7(1), 20220158. https://doi.org/10.1515/opis-2022-0158

Wang, J. L., Li, R., Yang, J., & Mao, C. (2024). RAFT: Realistic attacks to fool text detectors. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. https://aclanthology.org/2024.emnlp-main.939/

Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., Afzal, O. M., Mahmoud, T., Puccetti, G., Arnold, T., Aji, A. F., Habash, N., Gurevych, I., & Nakov, P. (2024). M4GT-Bench: Evaluation benchmark for black-box machine-generated text detection. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). https://aclanthology.org/2024.acl-long.218/

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. https://doi.org/10.1007/s40979-023-00146-z

Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., & Wong, D. F. (2025). A survey on LLM-generated text detection: Necessity, methods, and future directions. Computational Linguistics, 51(1), 275-338. https://aclanthology.org/2025.cl-1.8/

Xu, Y. E., Zhong, Z., Raghunathan, A., Fang, F., & Kolter, J. Z. (2026). Base models look human to AI detectors. arXiv. https://arxiv.org/abs/2605.19516

Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P., Soraperra, I., & Rahwan, I. (2025). Empirical evidence of large language model's influence on human spoken communication. arXiv. https://arxiv.org/abs/2409.01754

Yang, X., Pan, L., Zhao, X., Chen, H., Petzold, L., Wang, W. Y., & Cheng, W. (2024). A survey on detection of LLMs-generated content. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 9786-9805). https://aclanthology.org/2024.findings-emnlp.572/

Zhang, S., & Lv, X. (2026). AI detection risks undermining academic integrity. Nature Human Behaviour. https://doi.org/10.1038/s41562-026-02528-y

Zheng, J., et al. (2025). TH-Bench: Evaluating evading attacks via humanizing AI text on machine-generated text detectors. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (V.2). https://doi.org/10.1145/3711896.3737418

Zhou, Y., He, B., & Sun, L. (2024). Humanizing machine-generated content: Evading AI-text detection through adversarial attack. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 8427-8437). https://aclanthology.org/2024.lrec-main.739/