Sentence-Length Burstiness as a Cross-Disciplinary and Cross-Model Signal of AI Rewriting

TextPulse Research · Working Paper

Abstract

Burstiness, the uneven rhythm of sentence lengths in a text, is one of the most explicit differences between human and AI writing, and among the least precisely measured. This study measures it at scale. We compared 60,779 human-written academic texts with rewrites of the same texts generated by eight AI model configurations across six model families, so that every burstiness change is caused by the model rather than by topic or content. A text's burstiness is the standard deviation of its sentence lengths, and its coefficient of variation when overall length is factored out. Human academic writing in this corpus has a mean coefficient of variation of 0.449. The AI generated rewrites average 0.376, and 79.3 percent of all AI rewrites are flatter than their human written source. The flattening is universal across models but varies eight times in strength, from a mean change of 0.028 for the most rhythm-preserving configuration to 0.234 for the strongest flattener, which produced a flatter text in 98.5 percent of its rewrites. Models trim long sentences and remove the short, direct sentences human writers use, and converge on a uniform spectrum of medium-length sentences that reads monotonous and stiff. Disciplines differ as predicted, with humanities and social science prose longer and burstier than engineering prose, and this variation makes any single burstiness threshold misfire unevenly. A threshold that catches 62 percent of rewrites also flags 39 percent of genuine human texts, and flags science authors most. Burstiness is a strong population-level signal, but an unsafe individual-level verdict. Per-text sentence statistics and all code are made publicly available.

1. Introduction

An experienced reader can spot machine writing by describing its 'rhythm' before its vocabulary. Human writing is active and bursty. A short, direct sentence explains a point. A long sentence then describes the qualifications, and the alternation between the two gives writing its 'pulse'. In contrast, machine writing is uniform. Every sentence carries roughly the same weight, at roughly the same length, in the same register, and the effect over a paragraph is monotonous and flat.

This difference has a name, burstiness, and it appears in most public explanations of how AI detectors work. The term itself is a source of confusion. It is used for at least three different measurements, the original detector definition involves model perplexity rather than sentences, and the company that popularized the term retired it from production in 2023. Meanwhile the

simple, testable version of the claim, that AI models write with less sentence-length variation than humans, has never been tested with a design that controls for content.

In this study, using 60,779 paired texts, each a human academic text and a model's rewrite of that same text, we measure sentence-length variation on both sides and attribute the difference to the model. Comparing arbitrary human and AI corpora confuses topic, genre, and length with authorship. Comparing an AI rewrite with its own human written source isolates what the model did to the author's rhythm. We ask four primary questions. How large is the flattening effect overall? Does every model flatten, and by how much? How does burstiness vary across academic disciplines? What do the answers imply for burstiness as a detection signal? Before the results, Section 2 defines the measure in its simplest form and explains how AI detectors have used burstiness in the scores they report.

2. What burstiness means

2.1 Three senses of one word

The term must be clarified before we present the methods. In the detection industry's original usage, burstiness was defined against model probability. GPTZero, which popularized the term in early 2023, defined it as "a measure of how much writing patterns and text perplexities vary over the entire document", where perplexity measures how unpredictable a text is to a language model. By that definition burstiness is the variance of per-sentence perplexity, and not a property of sentence lengths. GPTZero states that it stopped using perplexity and burstiness for detection in autumn 2023 when it moved to a classifier-based architecture. In corpus linguistics, burstiness carries an older and different meaning. It describes how occurrences of a word cluster in time or position rather than arriving evenly, a property of content words known since the statistical language modeling of the 1990s.

The sense that spread into popular explanations, and the one this paper measures, is the simplest form. A text is 'bursty' when its sentence lengths vary widely, and 'flat' when they do not. This version needs no language model, can be computed by hand, and is the property readers perceive as 'rhythm'. Section 2.2 states it formally.

2.2 The measure, in its simplest form

Burstiness here is *the standard deviation of sentence lengths*. As a simplified example, consider a five-sentence text with sentence lengths of 6, 21, 9, 30, and 14 words. The mean is 16 words. The deviations from the mean are -10, +5, -7, +14, and -2. Squaring them gives 100, 25, 49, 196, and 4, which average to 74.8, and the square root of that average is the standard deviation, 8.6 words. Now consider a rewrite of the same text with sentence lengths of 15, 17, 16, 18, and 14 words. The mean is again 16, but the deviations are -1, +1, 0, +2, and -2, the average of the squares is 2, and the

standard deviation is 1.4 words. Both texts have identical average sentence length. One is bursty while one is flat, and the standard deviation separates them by a factor of six.

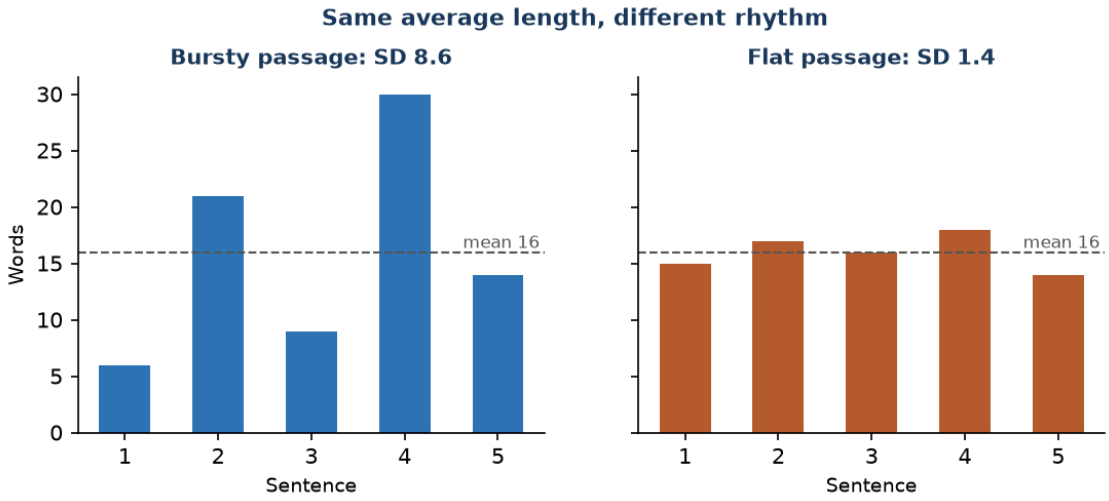

Figure F5. Two texts with the same mean sentence length and very different rhythm.

Since longer documents tend to have larger absolute deviations, we also report the coefficient of variation, CV, which is the standard deviation divided by the mean. The flat text above has CV 1.4 divided by 16, about 0.09. The bursty one has CV 0.54. CV lets texts of different overall length be compared on one uniform scale, and it is the primary statistic in the results obtained in this study. A text with long but uniform sentences still scores flat on CV, which is the property the measure is intended to capture.

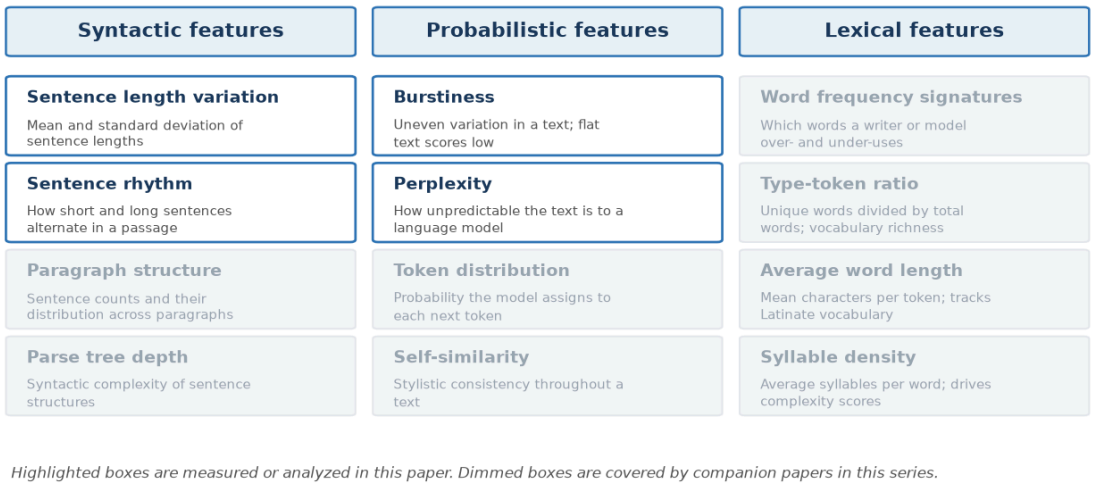

Figure F1. The feature families of AI text analysis. Highlighted families are measured in this paper.

2.3 How detectors turn burstiness into an AI score

An AI detector returns a score, usually a number between 0 and 100, that expresses how strongly a text resembles machine writing (i.e., AI generated text). Burstiness is a variable in that calculation in two ways. In the first generation of public tools it was used directly, as a threshold statistic. GPTZero's original foundation system computed two document-level numbers, the overall perplexity of the text and its burstiness, meaning the variation in per-sentence perplexity, and compared both against cutoffs calibrated on samples of human and machine writing (GPTZero, 2023). A text that was easy to predict and even from sentence to sentence was considered to be on the machine side of the cutoffs. This formula's weakness is the overlap problem this paper measures in Section 5.5, and GPTZero discarded both statistics from its production system in autumn 2023 (GPTZero, 2024).

The second use is more durable, as one variable inside a trained classifier. Feature-based detectors represent each text as a list of measurable properties, for example the average sentence length, the standard deviation of sentence length, punctuation counts, and vocabulary richness, and feed that list to a standard machine learning model such as logistic regression or a random forest. The model is trained on labeled human and machine texts and learns how much weight to give each variable. For a new text it outputs a probability, and that probability is what the user sees as the AI score. Fröhling and Zubiaga (2021) built classifiers of this kind against GPT-2, GPT-3, and Grover and found them competitive with more expensive detection methods. Desaire et al. (2023) separated ChatGPT chemistry writing from human academic writing with a 20-feature classifier in which the variation in sentence length ranked among the most useful variables. Kumarage et al. (2023) showed that adding stylometric features of this kind, including sentence-length statistics, improves transformer-based detectors on shorter texts. In this design burstiness is never a verdict on its own. A flat rhythm raises the score gradually, and the other variables can outweigh it.

Most current commercial detectors use a third design, an end-to-end neural classifier that reads the text directly and learns its own internal features (Crothers et al., 2023). Burstiness is no longer an explicit variable there, but the network can learn the same structure implicitly, which is why flat, uniform prose still tends to receive high AI scores from tools that never compute a sentence-length statistic. Regardless of the design, the information a detector can extract from burstiness is limited to how well the measure separates human from machine text, which is the primary investigation in this study.

3. Literature review

Sentence-length variation is one of the oldest measures in linguistic stylometry. In early foundation work, Yule (1939) treated sentence-length distributions as a statistical fingerprint of authorship in disputed-attribution cases. Williams (1940) showed that these distributions differ measurably and stably between authors. The measure has since remained in use for authorship analysis. The corpus-linguistics sense of burstiness, the uneven clustering of word occurrences, was formalized in statistical language modeling by Church and Gale (1995), and then extended to content words and phrases by Katz (1996).

In AI detection, probability-based signals originated first. GLTR visualized how consistently machine text draws from a model's high-probability tokens (Gehrmann et al., 2019). DetectGPT showed that model-generated text occupies distinctive regions of the model's probability function (Mitchell et al., 2023). GPTZero introduced the terms perplexity and burstiness to the public in 2023, then retired both from its production system later that year, which is documented in its own support materials. The sentence-length version of burstiness remained common in explainers and classroom advice, without direct measurement. In parallel, a feature-based line of work preserved sentence statistics inside detection itself, as classifier inputs rather than public explanations (Fröhling & Zubiaga, 2021; Kumarage et al., 2023), and the general detection literature is surveyed by Crothers et al. (2023).

Direct measurements of the sentence-length claim have started to appear for generated text. Muñoz-Ortiz et al. (2024) compared human news writing with the output of six language models across morphological, syntactic, and psychometric dimensions and found that the human texts show more scattered sentence-length distributions. Herbold et al. (2023) found that ChatGPT essays carry a consistent linguistic fingerprint that distinguishes them from human student essays. Desaire et al. (2023) found sentence-length variation to be among the most informative variables for separating machine from human science writing. These studies compare separate human and machine corpora, so topic, genre, and length vary together with authorship. But they do not preserve the content itself as a constant, and none measures the AI rewriting case, where a model restructures a text a human already wrote, which is the primary motivation behind this study.

Related evidence involves homogenization. Our vocabulary study of this same corpus found that models rewrite academic text into a small lexical register dominated by formal connectives and Latinate substitutions (TextPulse Research, 2026d), and word-frequency studies of the published literature document the same register at population scale (Kobak et al., 2025; Liang et al., 2024). Rhythm is the structural counterpart of that lexical register, and it has not been measured with the content itself preserved as a constant, the main aim of this research. Paraphrasing collapses the accuracy of probability-based detectors (Krishna et al., 2023), and our detector agreement study found that commercial detectors disagree sharply on identical academic texts (TextPulse Research, 2026a). A signal this heavily cited should have its effect size and its false-positive behavior thoroughly investigated against a human-AI parallel corpus.

4. Data and methods

4.1 Corpus

The audit uses the corpus of 60,786 human-AI paraphrase pairs built from open-access academic text that served our citation fidelity and vocabulary studies (TextPulse Research, 2026c, 2026d). Each pair has a source text of roughly 100 to 400 words from human written scholarly writing and an AI generated rewrite by one of eight model configurations. Corpus 1 contains 38,323 pairs from two configurations under a prompt designed for faithful rewriting. Corpus 2 contains 22,463 pairs

from six configurations in the DeepSeek, Grok, Mistral, Qwen, Gemini, and OpenAI families under a common rephrase prompt. Discipline labels from both corpora are normalized into nine groups. The corpus was reduced to 60,779 pairs after discarding short texts under three sentences.

4.2 Measures and statistics

Sentences are split with a rule-based segmenter applied identically to both sides, and sentence length is counted in words. For each text we compute the mean sentence length, the standard deviation, and the CV. The paired outcome is the change in CV from the human source to the AI rewrite. We report means with 95 percent confidence intervals, the share of pairs whose rewrite is flatter than its source, per-model and per-discipline breakdowns, and a threshold analysis that treats CV as a naive detector and computes for each cutoff the share of rewrites caught and the share of genuine human texts wrongly flagged. All code and the per-text statistics are released.

5. Results

5.1 Rewriting flattens sentence rhythm

Human academic texts in this corpus have a mean CV of 0.449. Their AI rewrites have a mean of 0.376. The mean paired change is -0.073 (95 percent confidence interval -0.074 to -0.072), and 79.3 percent of all 60,779 rewrites are flatter than their source. The average source text has a sentence-length standard deviation about 12.0 words, and the average rewrite about 9.4, a reduction of 2.6 words of spread per text.

The flattening comes from both ends of the distribution. Models generally trim the longest sentences, and remove the short, direct sentences almost entirely. Mean sentence length falls slightly in every discipline, so the rewrites are not longer on average, but they are more uniform. The characteristic human pattern, a short direct sentence followed by a long analytical one, becomes an unbroken series of medium-length sentences in an AI generated rewrite. The prose reads monotonous and stiff because the variation between short and long sentences was decreased.

5.2 An example pair

Figure F2 shows the effect in a single real pair from the corpus, an economics text rewritten by DeepSeek. The human source comprises twelve sentences with lengths from 6 to 45 words, a bursty, direct rhythm with a standard deviation of 9.7. The rewrite also runs twelve sentences, all between 14 and 21 words, uniform and flat, with a standard deviation of 2.1. The two versions state the same content. Their rhythm is entirely different.

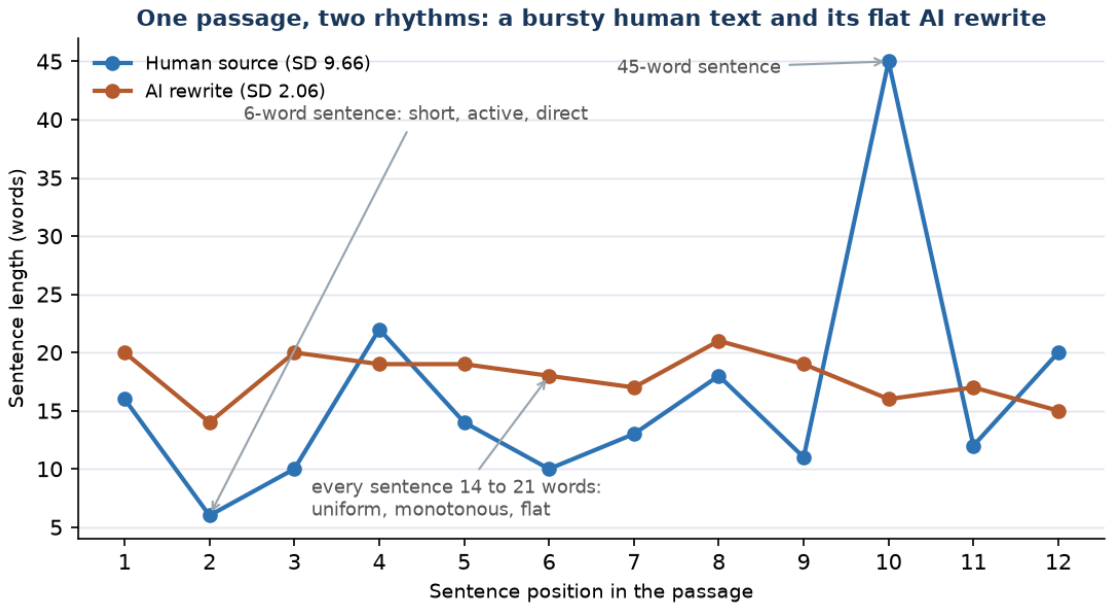

Figure F2. Sentence lengths across one text and its rewrite. The human line is bursty and direct. The AI line is uniform, monotonous, and flat.

5.3 Every AI model flattens

All eight models flatten on average, and the strength varies eight times between the weakest and the strongest. Within corpus 2, where six families rewrote under the same prompt, the DeepSeek configuration is the strongest flattener, with a mean CV change of -0.234 and a flatter rewrite in 98.5 percent of its pairs. OpenAI follows at -0.123, then Gemini, Mistral, and Qwen near -0.10, while Grok preserves rhythm best at -0.032. The two corpus 1 models, produced under a prompt built for faithful rewriting, change CV by only -0.028 and -0.063, which shows that rhythm preservation is generally trainable based on prompt.

The ordering does not match the vocabulary results on the same corpus. Grok, the strongest vocabulary rewriter in our companion study, changes rhythm the least, and DeepSeek combines mid-range vocabulary change with the strongest flattening. Style fidelity is not one property. A model can preserve an author’s words while removing their rhythm, or vice versa.

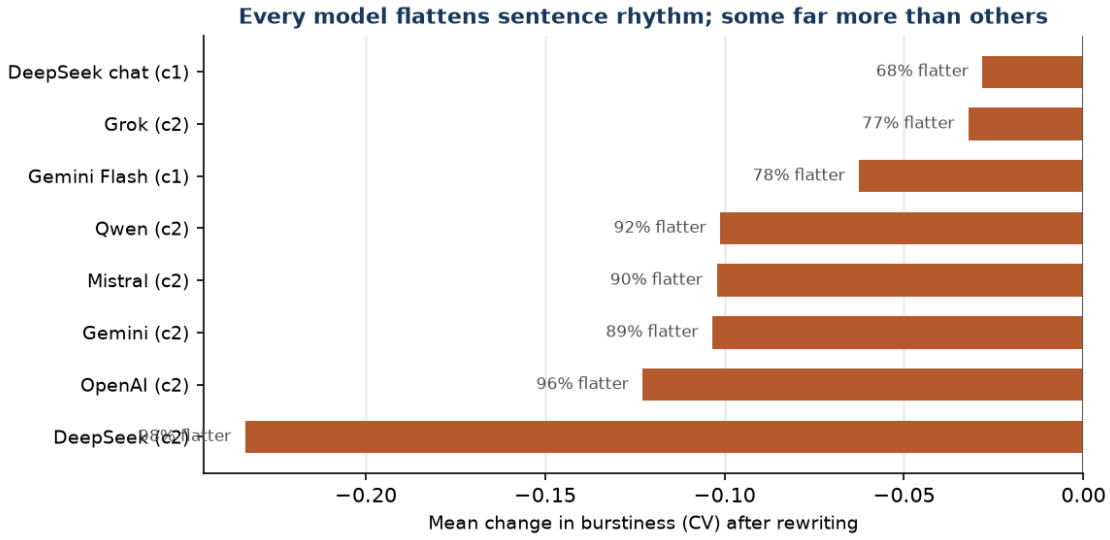

Figure F3. Mean change in burstiness by model. Labels give the share of each model's rewrites that were flatter than their sources.

5.4 Disciplines write differently, and the models flatten them all

Human writing styles differ by field in the expected direction. Humanities and law writing has the longest and burstiest sentences, averaging 29.9 words with a CV of 0.484. Social science writing averages 27.2 words, wordier and more varied than engineering and computer science writing at 24.3 words, with natural sciences similarly compact and even. The gap between the extremes is more than five words of average sentence length, which is a visible stylistic difference, and rewriting compresses every field toward the same flat profile.

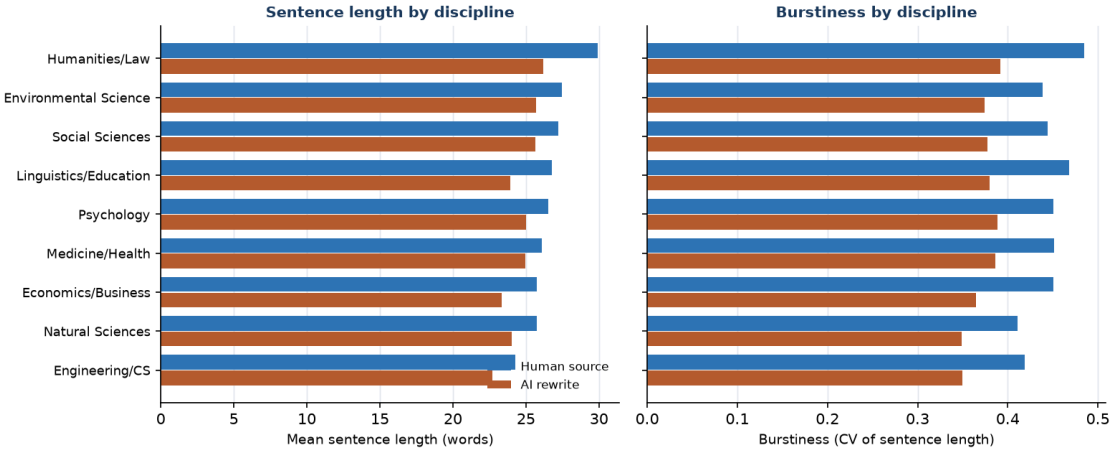

Figure F4. Sentence length and burstiness by discipline, human sources against AI rewrites.

5.5 The signal is strong for populations

Treating CV as a naive one-number detector makes the risk concrete. A cutoff of 0.40 catches 62 percent of rewrites in this corpus, but it also flags 39 percent of the genuine human texts. Tightening the cutoff to 0.30 drops the false flags to 10 percent and the catch rate to 28 percent. No cutoff separates the populations effectively, since bursty and flat human writers both exist, and the false flags are distributed unevenly. At the 0.40 cutoff, 52 percent of natural science texts and 49 percent of engineering texts written by humans are below the line, against 31 percent for humanities and law. A burstiness rule that ignores discipline penalizes the fields whose ordinary style is compact and even. Burstiness distinguishes distributions with high confidence. It does not identify individuals, and our detector agreement study shows what happens when tools overlook that distinction (TextPulse Research, 2026a).

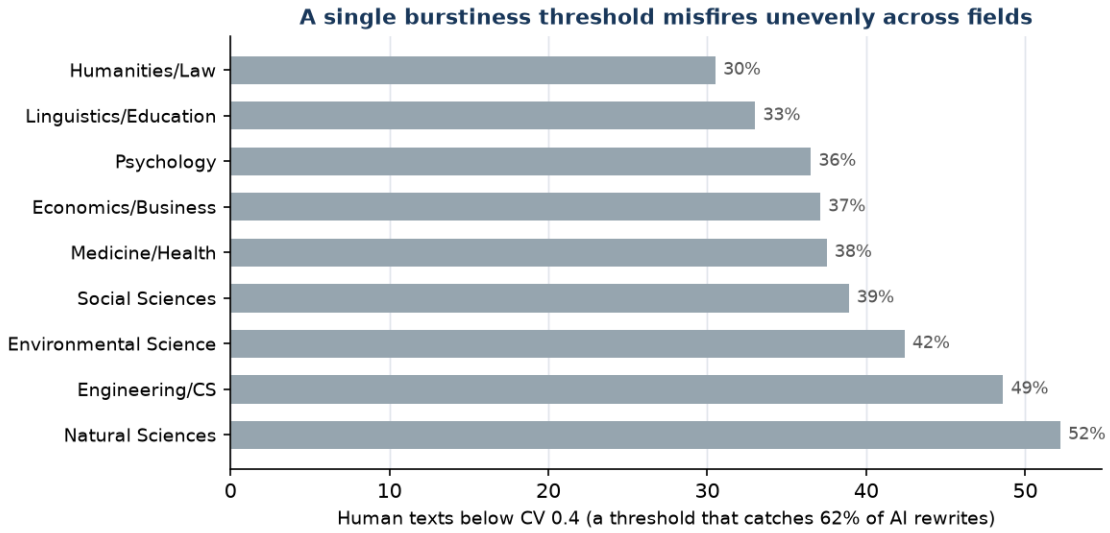

Figure F6. Share of genuine human texts flagged by a burstiness threshold that catches 62 percent of rewrites, by discipline.

6. Discussion

AI models write with less sentence-length variation than the human texts they rewrite. This effect appears in four of every five rewrites, and it is caused by the model rather than by the content, since the content is fixed by design. The mechanism is consistent with how aligned models write more generally. Rewriting toward high-probability output pulls sentence structure toward typical lengths in the same manner it pulls word choice toward preferred vocabulary, and the two effects together produce the register readers describe as flat, uniform, and machine-like. Our vocabulary study measures the lexical half of that convergence, while this study measures the structural half.

For writers and editors, the practical signal is the missing short sentence. The clearest single difference between the human and rewritten texts is the absence of sentences under ten words, the ones a human uses to state a claim directly before qualifying it. Restoring variation, by

reintroducing short direct sentences, addresses the most measurable component of machine-sounding writing.

For detection, the numbers support using burstiness only as one signal among several, never as a sole verdict. The population gap is large, but the individual overlap is larger than the public discussion assumes, and the false flags concentrate on science and engineering authors whose natural style is compact. The feature-based classifier literature reaches the same conclusion, treating sentence-length variation as one weighted input among multiple rather than as a single decision rule (Fröhling & Zubiaga, 2021; Desaire et al., 2023). GPTZero's own retirement of burstiness from production in 2023 is in line with this reading. A signal can be explanatory, but still insufficient on its own.

This study has several limitations. Sentence segmentation is rule-based, and although it is applied identically to both sides of every pair, unusual formatting can miscount sentences. The corpus is English academic writing only, and rhythm norms differ in other genres and languages. The two corpora used different prompts, so cross-model comparison is limited to corpus 2, and the corpus 1 results show that prompt effects are significant. Burstiness here is sentence-length variation only. The perplexity-variance and word-clustering senses of the term are related but distinct measurements, and results for one do not automatically transfer to the others.

7. Conclusion

Rewriting by AI language models reduces sentence-length variation, and the reduction is large and consistent. Across 60,779 paired texts, AI rewrites lost a sixth of their sentence-length variation on average, four in five were flatter than their sources, the strongest model flattened nearly every text it rewrote, and the effect held in every academic discipline. The flattening is measurable with simple standard deviation, visible in a single figure, and strong enough to characterize populations of text with high confidence. The overlap between individual authors is too large to justify a verdict on any single document, and the false flags concentrate on the disciplines whose human style is most compact and even (e.g., engineering). Burstiness is a valid population-level descriptor of machine prose, but not a valid basis for a final verdict on any individual text alone.

Data availability

Per-text sentence statistics for all 121,558 text sides (pair identifier, corpus, model, discipline, side, word and sentence counts, mean, standard deviation, and CV of sentence length), the example-pair sentence lengths, and all analysis and figure code are openly available on Zenodo at https://doi.org/10.5281/zenodo.22033062.

References

Church, K. W., & Gale, W. A. (1995). Poisson mixtures. Natural Language Engineering, 1(2), 163-190.

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

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

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

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, 111-116. arXiv:1906.04043.

GPTZero. (2023). Perplexity and burstiness: What is it? https://gptzero.me/news/perplexity-and-burstiness-what-is-it/

GPTZero. (2024). How do I interpret burstiness or perplexity? GPTZero Support Center. https://support.gptzero.me/articles/9585228410-how-do-i-interpret-burstiness-or-perplexity

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. Scientific Reports, 13, 18617. https://doi.org/10.1038/s41598-023-45644-9

Katz, S. M. (1996). Distribution of content words and phrases in text and language modelling. Natural Language Engineering, 2(1), 15-59. https://doi.org/10.1017/S1351324996001246

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

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

Kumarage, T., Garland, J., Bhattacharjee, A., Trapeznikov, K., Ruston, S., & Liu, H. (2023). Stylometric detection of AI-generated text in Twitter timelines. arXiv:2303.03697.

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 (ICML 2024), PMLR 235. arXiv:2403.07183.

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 (ICML 2023). arXiv:2301.11305.

Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting linguistic patterns in human and LLM-generated news text. Artificial Intelligence Review, 57. https://doi.org/10.1007/s10462-024-10903-2

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. (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

Williams, C. B. (1940). A note on the statistical analysis of sentence-length as a criterion of literary style. Biometrika, 31(3/4), 356-361.

Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. Biometrika, 30(3/4), 363-390.