// WHITE PAPER · PUBLISHED 6 AUGUST 2026

The Flat Line

Finding machine-written prose by eye, and what the finding is worth once you have it

Jason Newell · MAX Research Collective · Report WPR-AGT-002 · Published 6 August 2026
Licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0),
https://creativecommons.org/licenses/by/4.0/
Original: https://www.maxresearchcollective.com/articles/the-flat-line.html
Cite as: Newell, J. (2026). The Flat Line. MAX Research Collective, WPR-AGT-002.
https://www.maxresearchcollective.com/articles/the-flat-line.html

In 1985 the assignment sheet came off a ditto machine and smelled of solvent. It asked for five paragraphs. Each one opened on a topic sentence and carried three supporting details. Then it handed off to the next. The last paragraph restated the thesis in different words. A comma splice cost half a letter grade.

That was writing instruction, and it worked the way piano scales work. You learned the pattern. You reproduced the pattern. Reproduction was the evidence that the lesson had landed. Reading got assigned for the same reason. A student who reads constantly absorbs the shapes that sentences come in, the way a chess player absorbs openings years before being able to name any of them.

Forty years on, that trained evenness is what gets a paper flagged as machine-written.

The complaint arrived as two words. "AI slop" now covers the spam listicle, the vendor blog, the status report from the desk across the aisle, and the cover letter that reads like every other cover letter in the stack. The accusation gets made constantly. It almost never carries a reason. Something about the page felt off, so the reader filed it under machine and moved on.

The question the Max Research Collective set out to answer is narrower than the complaint. Can a person identify machine-written prose by reading it, with no detector, no scanner, and no statistics package open in the next window?

Mostly they can. Confidence collapses exactly where the cost of being wrong runs highest, and that second half is the part worth carrying around.

Every draft in this shop gets swept against a written catalog of machine tells before it ships, and this one was swept too. That catalog exists because its patterns keep surfacing in work no machine touched. There is the difficulty, in one sentence.

What the machine learned from everybody

A large language model predicts the next stretch of text from everything in front of it. Train it on an enormous body of existing writing and it learns which continuation is likeliest. Likeliest does the work in that sentence. Averaged over billions of examples, the likeliest continuation sits in the middle of the range. So the model writes toward the middle by construction, on every dimension at once.

There is a technical name for the surprise involved. Perplexity measures how unexpected each next word is to a predictor that has read everything before it. Low perplexity means text a machine finds easy to guess. Burstiness is the companion measure. It tracks how far that predictability swings from one sentence to the next.

Human writing swings. A person reaches for an odd word, drops a private reference nobody else would use, writes four words because the beat ended there, then lets the next sentence run thirty-eight words because the thought genuinely took that long to finish and cutting it in half would have broken it. A model has no bad night, no deadline, and no obsession. Nothing in it wants to overshoot.

So phrase-hunting fails. Scrub ”delve” and ”tapestry” out of a draft and nothing about its shape changes. Shape is what a careful reader registers first, usually a paragraph before the vocabulary registers at all.

The property that gives it away

Uniformity is the signal. Sentence lengths converge on one value. Paragraphs settle into one size. Every section opens with the same move and closes on the same kind of line, and the register holds one temperature from the first sentence to the last.

Figure 1 charts sentence lengths in reading order for two samples. Above is a published field report from this catalog, 188 sentences of prose written before anybody measured it. Below is a machine draft on a related subject, produced on request in default register and left unedited.

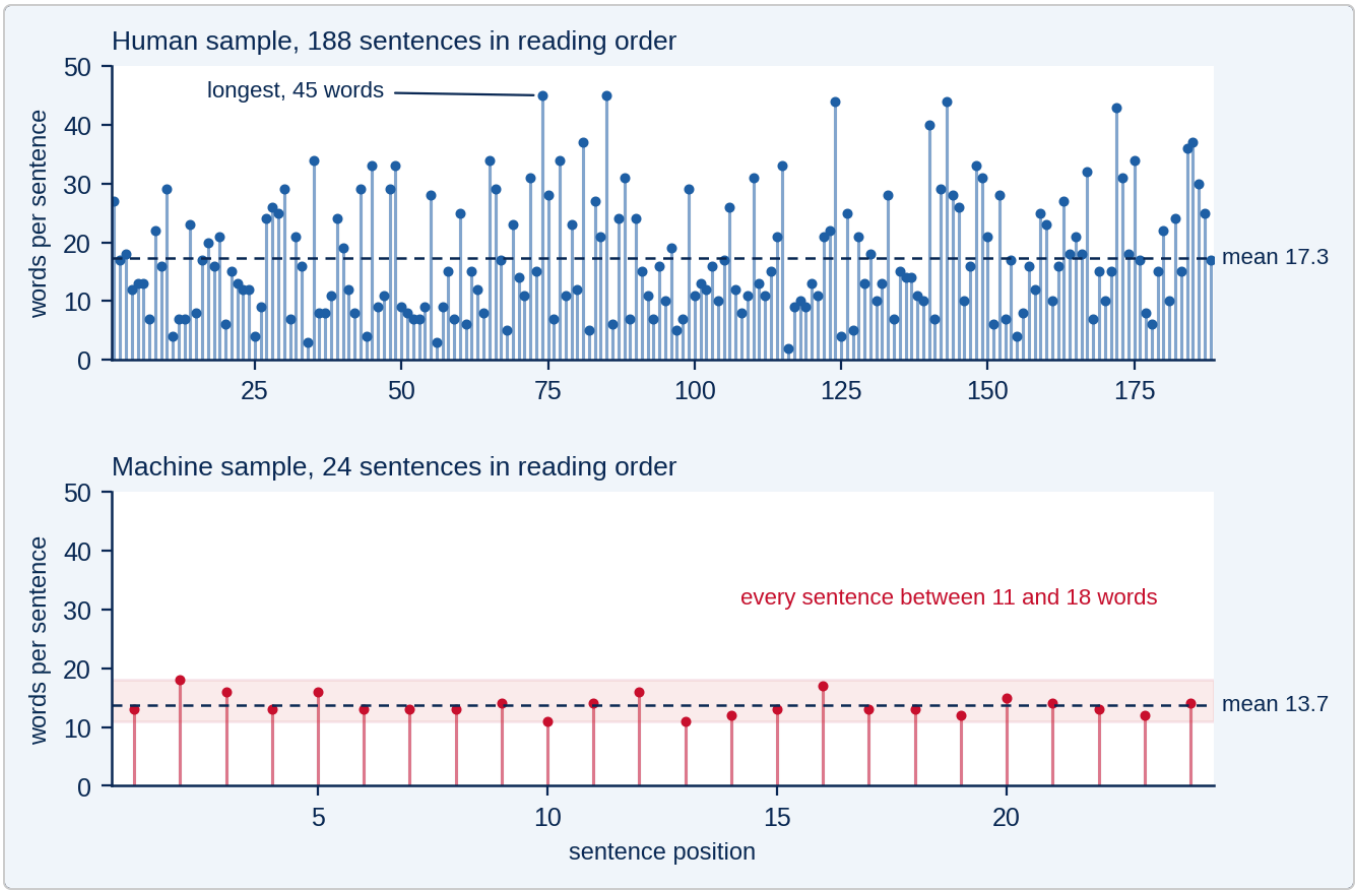


Figure 1. Look at the vertical spread rather than the height. The human series reaches 45 words and drops to 2. The machine series never leaves the band between 11 and 18.

Numbers behind the picture separate the two samples cleanly, and they do it on one axis only.

MEASURE	HUMAN SAMPLE	MACHINE SAMPLE
Sentences counted	188	24
Mean sentence length	17.3 words	13.7 words
Standard deviation	9.9 words	1.8 words
Coefficient of variation	0.57	0.13
Shortest sentence	2 words	11 words
Longest sentence	45 words	18 words
Sentences of 30 words or more	12.8 percent	0 percent
Paragraph-length variation	0.34	0.05

Table 1. The averages nearly agree. Every measure of spread disagrees by a factor of four or more.

Read the first two rows, then the third. Mean sentence length puts the samples within four words of each other. A reader checking averages would learn nothing at all. Standard deviation is roughly the average distance between one sentence and the average sentence, and it runs 9.9 words in the human sample against 1.8 in the machine one. Divide each by its own mean and the comparison survives across writers with different habits: 0.57 against 0.13, better than four times the spread.

No sentence in the machine sample runs under 11 words or over 18. Every idea in it arrived in a container of the same size. That is the flat line. Once a reader has seen it charted, they start seeing it on the page without the chart.

Seven checks and a pen

None of this needs software. Each check takes under a minute on a page of prose, and they run in the order that finds the most with the least reading.

#	CHECK	WHAT TO DO	MACHINE DRAFT	HUMAN DRAFT
1	The lurch	Count the words in ten consecutive sentences, writing the ten numbers in the margin	Numbers cluster, often within five of each other	Numbers jump, with real short beats and at least one long run
2	The opener column	Read down the first word of every sentence in one section	The same handful of words repeat, and three in a row often match	Entry points vary, on clauses, objects, and times
3	The squint	Hold the page at arm's length and look only at the block shapes	Even bricks, or a tidy long-short alternation	Uneven blocks in no predictable sequence
4	The last lines	Read only the final sentence of each section, one after another	Short, quotable, each one resolving its section into a lesson	Some end on an image, a number, or mid-gesture
5	The explaining sentence	Find any sentence whose only job is to say what the sentence before it meant	Present throughout, often once a paragraph	Rare, because the concrete sentence was trusted to work
6	The attention map	Mark how much room each point got	Every point gets roughly equal space	One point gets three paragraphs and another gets a clause
7	The checkable detail	Pick the most specific claim on the page and try to verify it	Plausible, decorative, unsourceable	Traceable, and sometimes inconveniently so

Table 2. The pass runs in two stages. Checks 1 to 3 read the shape of the page, and checks 4 to 7 read the writer's decisions.

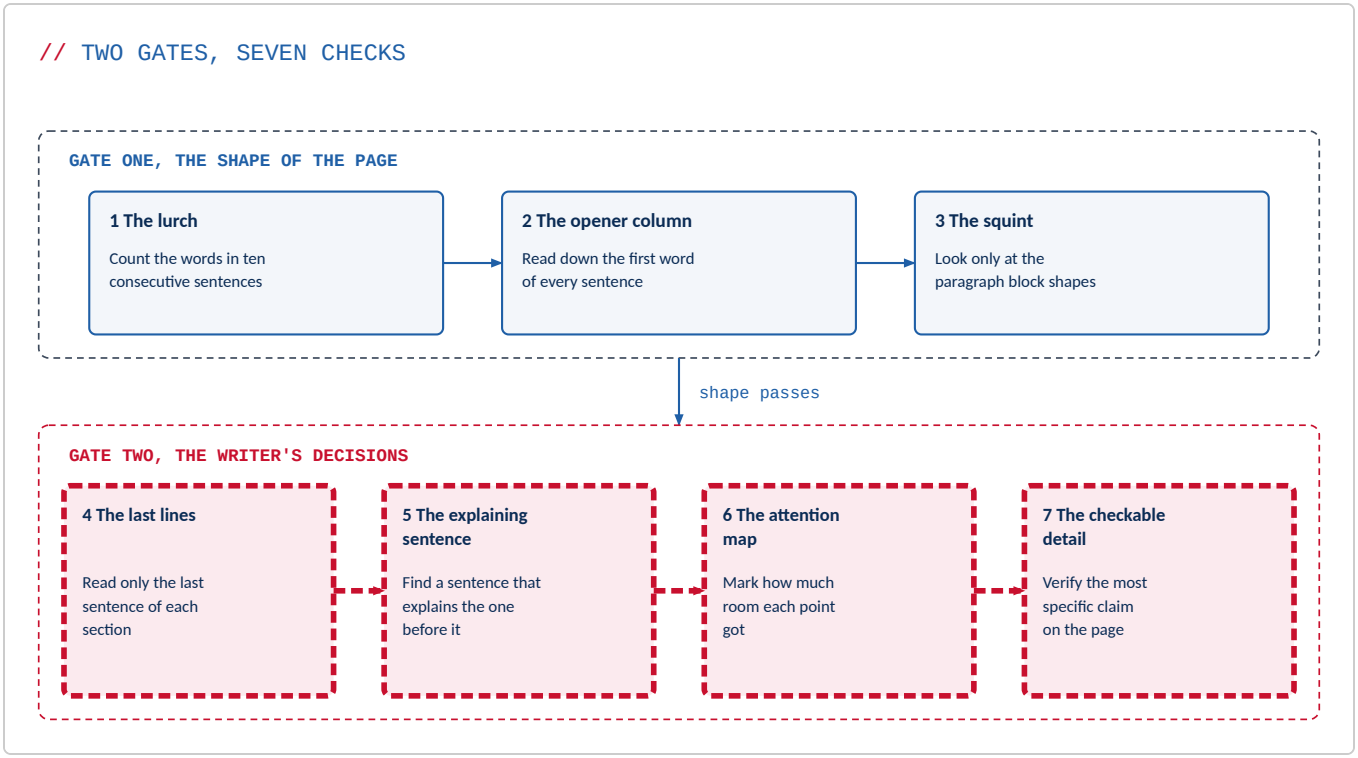


Figure 2. Two gates, seven checks. A draft that clears the shape gate and then the decision gate was written by somebody who was choosing.

Check one carries most of the weight, and people skip it because counting feels beneath them. It is not beneath them. Ten numbers in a margin settle more arguments about a page than an hour of reading it for vocabulary.

The opener column is the fastest of the seven. Cover the page so only the left edge shows, then read down. A rut there is one the writer never noticed; a reader feels it well before naming it.

Check five earns its own paragraph, because automated scanners miss it almost entirely. A model writes a concrete sentence. Then, distrusting the reader, it writes a second sentence explaining what the first one meant. In a story that looks like a gesture followed by its own interpretation. In technical prose it looks like a measurement followed by a line announcing what the measurement proves. Cut the second sentence of each pair. A machine draft loses about a tenth of its length and nothing else.

Number six is the check a good writer fails on purpose. Human writers dwell where they care and skate where they don't. Lumpy proportions come out of that, and no assignment sheet would accept them. Even coverage is what a system produces when it has no preferences.

What the checks are actually counting

Each check above is a formal measure worn as a habit. The measures matter because they can be argued about. A reader who knows what the eye is approximating can also tell when the eye is wrong.

FORMAL MEASURE	COMPUTATION, IN PLAIN SPEECH	WHAT IT DETECTS	HOUSE BAND
Coefficient of variation	Standard deviation of sentence lengths divided by the mean	Whether sentences vary or converge	0.65 to 0.95
Short-sentence share	Sentences of five words or fewer, over all sentences	Whether the short end of the range is populated	8 to 18 percent
Long-sentence share	Sentences of forty words or more, over all sentences	Whether any single idea was allowed to run	2 to 6 percent
Paragraph-length variation	The same ratio applied to words per paragraph	Whether the page has one block shape or many	0.70 to 1.00
Opener concentration	Share of sentences beginning with the five commonest first words	A rut in sentence entry points	30 to 45 percent
Repeated five-gram share	Share of five-word runs that already appeared earlier	Restatement across sections	Under 1.5 percent
Moving-average type-token ratio	Distinct words in each 400-word window, over 400	Vocabulary range, with the caveat below	Above 0.40
Rule-of-three density	Three-item lists per 500 words	The tricolon used as a default rhythm	One or fewer

Table 3. The measures behind the seven checks, with the bands this shop writes toward. One row outside its band means nothing on its own.

Two rows need a warning printed beside them. Type-token ratio drops when a writer keeps the vocabulary deliberately plain. That is a craft decision. The row is ambiguous by nature, and it gets confirmed by reading or not at all. Coefficient of variation is worse, because it invites gaming. Chop good sentences in half to raise

the number and out comes a short-long-short-long waveform, which is itself among the loudest machine signatures on the list. The ratio is a diagnostic. Treating it as a target manufactures the defect it was built to catch.

One measure failed on the samples in Figure 1, and reporting it beats quietly dropping it. Opener concentration came in at 42.6 percent for the human sample against 45.8 for the machine one. The check separated nothing, and both figures sit near the top of the same house band. Bear in mind that the human sample was nearly eight times larger, which alone could account for the gap. A measure that discriminates on one pair of documents and not on another has a confidence interval, and none of these carry one.

Where the reading fails

People who build detectors know this ground better than people who buy them. OpenAI released a classifier for machine-written text in January 2023 and published its own numbers alongside it. It correctly identified 26 percent of AI-written text. It labeled human-written text as AI-written 9 percent of the time. One in four caught, one in eleven innocent pages accused. The tool came down on July 20, 2023, and the company cited its low rate of accuracy.

Bias in that error rate is not random. A 2023 study in Patterns ran 91 TOEFL essays by non-native English speakers and 88 essays by United States eighth-graders through seven widely used detectors. The detectors handled the American eighth-grade essays accurately. They misclassified more than half of the TOEFL essays as machine-written, at an average false-positive rate of 61.3 percent. At least one detector flagged 97.8 percent of the non-native essays. Essays that every detector agreed on shared one property: unusually low perplexity, meaning word choices a predictor found easy to guess.

Read that mechanism again. It is the same mechanism the seven checks run on. A writer working in a second language reaches for the common word, holds sentences to a similar length, and keeps the syntax safe. Out comes the statistical fingerprint of machine prose, for reasons that have nothing to do with machines.

Picture the sequence from inside that student's week: a paper handed in on time, a score returned by software nobody in the building can explain, an accusation with no author attached to it, and a grade held in abeyance while the burden of proof settles on the person with the least standing in the room.

// ONE THRESHOLD, TWO POPULATIONS

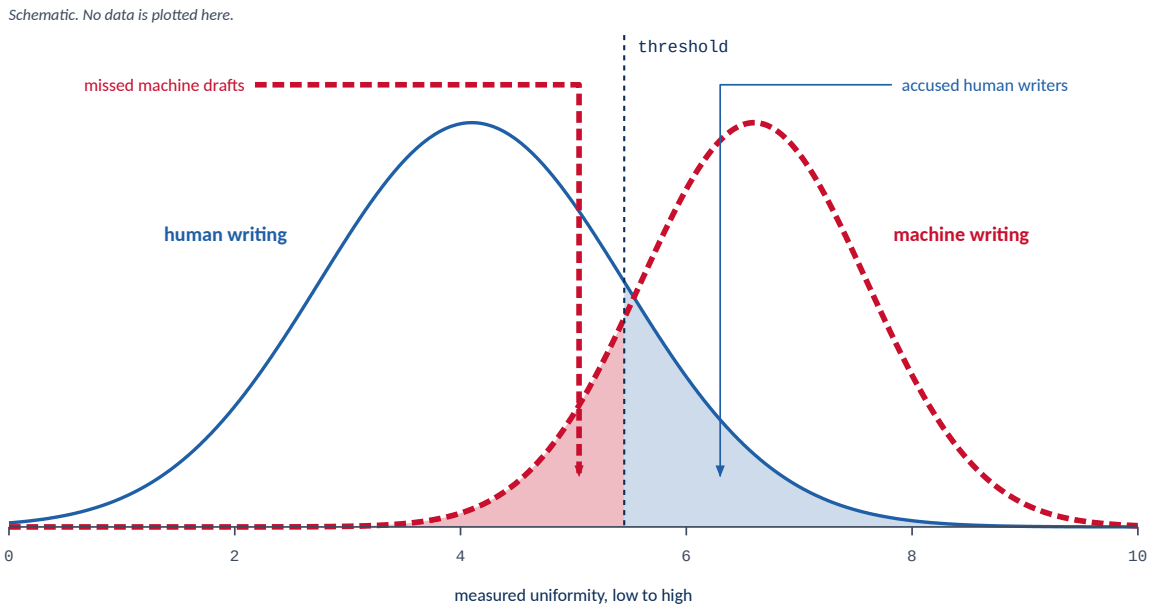


Figure 3. One threshold drawn through two overlapping populations. Machine texts left of the line are missed and human texts right of it are accused. Moving the line trades one error for the other.

WRITER	WHY THE PROSE READS MACHINE-MADE	WHAT THE CHECKS ARE ACTUALLY SEEING
Working in a second language	Common vocabulary, safe syntax, even sentence lengths	Low perplexity produced by caution
Technical and clinical documentation	House style enforces uniformity on purpose	A template working as intended
Student taught the five-paragraph form	The form specifies the uniformity	Successful compliance with instruction
Anyone who used a model to edit	Human argument, machine surface	A copyedit rather than an authorship
Heavily copyedited professional prose	House style sanded the idiosyncrasies off	The publisher's fingerprint over the writer's

Table 4. Five ways to look machine-made with no machine in the room. Row one carries the highest human cost. Row four is the one most policies have no language for.

Costs here are not symmetric, and any policy built on these checks has to say so out loud. A machine draft that slips past a reader costs that reader ten minutes. A false accusation costs a student a transcript, a grade appeal, and a relationship with a teacher. It lands hardest on the writer already working hardest to be understood.

So the honest position is narrow. Seven checks are good enough to decide what to read closely and what to set aside. They are good enough to reject a vendor white paper. They are nowhere near good enough to accuse a person. Most of the current damage sits in the gap between those two uses.

The classroom holds the sharpest version

Teachers are being asked to enforce a rule about AI while using AI to prepare the lesson that carries the rule.

Call it arithmetic instead of accusation. Gallup and the Walton Family Foundation surveyed 2,232 United States public school teachers in the spring of 2025. Sixty percent had used AI tools for work during that school year. About three in ten used them weekly, and those weekly users reported saving 5.9 hours a week, which comes to roughly six weeks over a school year. Lesson planning sat among the most common uses. Grading sat well down the list at 16 percent. Conversations with practicing teachers confirm the shape of that finding, more than a dozen of them, and not one was defensive about it.

Consider the position the rule puts a teacher in. Somebody covering five sections of thirty students, with essays due the same week, has a hundred and fifty documents to read and no reliable way of telling which came out of a chat window, which came out of a study group at somebody's kitchen table, and which came out of a kid who has always written like that.

Defense of the asymmetry is real and deserves stating properly. A lesson plan is an input. A student essay is the assessment itself, so a tool that speeds up the first can gut the second. The carpenter's nail gun doesn't excuse the apprentice from learning to swing a hammer, because the apprentice is being certified on a skill rather than delivering a house. Teachers under a workload that six recovered weeks materially relieves are not hypocrites for taking the relief.

The rejoinder is equally real. Students read the modeling before they read the rule. What the modeling says is that the tool is legitimate for the person with authority and illegitimate for the person without it. That is a lesson too, and it lands faster than any of the intended ones.

A policy that can't explain why the same tool changes moral character depending on who picks it up will get treated as an enforcement problem instead of an ethical one. Which is exactly how it is being treated.

Underneath all of it sits a question nobody has answered. Employers are moving toward AI-first workflows at speed, so schools are certifying a skill while the market repositions what the skill is for. Reading and writing still have to be learned, and the case for that never depended on employment. A person who can't write can't think in public. A person who can't read is at the mercy of whoever summarizes.

What survives when detection doesn't

Assessing an artifact is a losing position, and it gets worse every year. Assessing a process holds. Drafts with revision history, writing done in the room, five minutes of conversation about one paragraph: all of it survives contact with the tools, because the question moves from what the text is to what the writer can account for.

Ask a writer why they cut a paragraph. A person will tell you, usually at length, often with an opinion about the paragraph they kept instead. A generated draft has no answer there, because no choice was made, only a default taken. That one question outperforms every detector on the market and costs nothing to run.

Professional writing follows the same logic, and the accusation game there has no stakes and no point. Four properties keep turning up in prose a machine cannot currently produce on its own, and none of them is a phrase or a rhythm.

- Judgment. Somebody is weighing, and the sentence carries a verdict. "That gap came out of our own configuration" has a person behind it and no first-person pronoun in it.
- Stake. The cost lands on someone specific. A neutral report says coverage was reduced. A person says what the reduction cost.
- Selection. Uneven attention is a signature. What gets three paragraphs and what gets a clause is a decision no averaging process would make.
- Admitted error. Strongest of the four. "The case was built too early, and that misread the data" is a sentence no neutral document has ever contained, because nobody but the author would have written the mistake down at all.

A model can imitate all four on request. A writer who has none of them cannot fake any of them for three thousand words. The tell is the same tell as always, which is whether anybody was actually deciding.

The five-paragraph essay is still on the board in a great many rooms. It teaches the shape a machine writes in, and then the same room grades that shape as evidence of a machine.

Glossary

TERM	DEFINITION
Burstiness	Variation in how predictable a text is from sentence to sentence. Human prose swings between the obvious and the surprising. Machine prose holds a steadier level, which is why burstiness is one of the properties detectors and readers both key on.
Coefficient of variation	Standard deviation divided by the mean. Expressing spread as a fraction of the average makes it comparable across writers who work at different sentence lengths.
Corpus	A body of text used to train or evaluate a model. Everything a language model knows about how sentences go came from a corpus, so its writing is the average of that corpus rather than the voice of anyone in it.
False positive	A test result that says yes when the truth is no. Here it means human writing labeled as machine-written, and it carries almost all of the human cost in this subject.
Five-gram	Any run of five consecutive words. Counting how often a five-gram reappears measures how much a document restates itself.
Large language model	A program trained to predict the next stretch of text given everything before it. Prediction toward the likeliest continuation is what produces the central-tendency writing this article is about.
Machine-generated text detector	Software that estimates whether a passage was written by a model, usually by measuring perplexity and related statistics. Published accuracy figures have been poor enough that several vendors withdrew their tools.
Perplexity	A measure of how surprised a predictor is by each next word. Low perplexity means easily guessed text. It rises with unusual vocabulary, which is why enriching word choice moves a text out of the machine-looking range whether or not a machine wrote it.

TERM	DEFINITION
Register	The emotional and formal temperature of prose. Machine drafts tend to hold one register start to finish, where a person's writing warms, cools, and occasionally forgets itself.
Self-annotation	A sentence whose only job is to explain the sentence before it. It is the most reliable hand-detectable tell in this catalog and the one automated scanners miss most often, since nothing about it is lexical.
Standard deviation	Roughly the average distance between one measurement and the average of all of them. Applied to sentence lengths it answers a single question: do the sentences in this document agree with each other about how long a sentence should be.
Tricolon	A three-item list, as in "fast, cheap, and reliable." Humans use it for emphasis and models use it as a default rhythm, so density rather than presence makes it a tell.
Type-token ratio	Distinct words divided by total words, usually computed over a moving window so document length doesn't distort it. It approximates vocabulary range and drops for deliberately plain writing, which makes it the most easily misread measure in the set.
Uniformity	Low variance across several dimensions at once: sentence length, paragraph shape, sentence openers, register, and section-closing move. It is the single property that most reliably separates machine prose from human prose, and all seven hand checks approximate it.

Appendix B: The samples behind Figure 1 and Table 1

Both samples were measured with the same script, which strips tables, code blocks, headings, bulleted lists, and horizontal rules before counting, so the numbers describe running prose rather than document furniture.

Human sample. Sixty prose paragraphs from "AI Review at Agent Velocity," a field report published in this catalog, totaling 3,261 words across 188 sentences. Every

paragraph in the sample was matched against the delivered document before being counted, so the tables, pricing blocks, bulleted findings, and display arithmetic that piece carries are excluded and the numbers describe running prose only. It was written before any of these measures were applied to it, and it fails several of the current house rules, which makes it an honest specimen rather than a flattering one.

Machine sample. Twenty-four sentences across five paragraphs, 329 words, generated on request in default register on the subject of AI detection in education, then left unedited. Nothing about it was written to fail the checks. It shows how a model writes when nobody has told it not to, which is the condition the checks are built to catch.

The size mismatch is real. The human sample is roughly ten times longer. Measures that depend on frequency counts, opener concentration in particular, are less stable on the smaller sample. Sentence-length measures are stable at both sizes, and the argument rests on those.

Appendix C: Sources

CLAIM	SOURCE
Classifier accuracy of 26 percent true positive and 9 percent false positive, withdrawn July 20, 2023	OpenAI, "New AI classifier for indicating AI-written text," January 31, 2023, with the withdrawal notice added to the same page
False-positive rate of 61.3 percent on 91 TOEFL essays, near-perfect accuracy on 88 United States eighth-grade essays	Liang, Yuksekgonul, Mao, Wu, and Zou, "GPT detectors are biased against non-native English writers," Patterns, July 2023
Teacher AI use at 60 percent, weekly use near 30 percent, 5.9 hours saved per week	Walton Family Foundation and Gallup, "Teaching for Tomorrow: Unlocking Six Weeks a Year With AI," June 2025, surveying 2,232 United States public school teachers between March 18 and April 11, 2025
Sentence-length and opener statistics in Table 1	Measured for this article from the two samples described in Appendix B, with the human sample verified paragraph by paragraph against the delivered document

AGENTIC AI WPR-AGT-002

// ABOUT THE AUTHOR

Jason Newell leads organizations and people in technology, and is a researcher, deep technologist and builder — from a CRAY to agentic AI. He founded MAX Research Collective to build AI for regulated industries that runs where its owner has control, with healthcare as its strongest focus. About Jason → Connect on LinkedIn →