No writing tell separates human from machine on its own: the best of 10 reaches AUC 0.80

40 engineering texts published before 2020 against 120 texts from 4 models, counted per 1,000 words

Dmitriy Semenkevich · ORCID 0009-0009-3013-4978

21 August 2026 · dimhold.by

ABSTRACT

Readers convict each other's prose on single surface features: an em dash, a list of 3, the word delve. The argument runs without a denominator, because nobody publishes how often those features appear in writing that predates language models. This study fixed a threshold before collecting anything and then counted. Corpus A is 40 engineering texts whose publication date is mechanically verifiable as earlier than 2020, 72,082 words from Paul Graham, Joel Spolsky, Brendan Gregg, Julia Evans and the RFC series, taken in index order through a gate that rejected 152 candidates. Corpus B is 120 texts, 30 from each of 4 Anthropic models, written on the titles of the corpus A essays at the same target length with tools isolated. 10 rules count matches per 1,000 words. A rule counted as a machine fingerprint only if its rate was at least 10 times higher in corpus B and its per document distributions separated the corpora with an AUC of 0.90 or better. No rule cleared it. The best reached AUC 0.80, bootstrap interval 0.70 to 0.88, at a ratio of 2.6. The em dash runs 3.5 times higher in the machine corpus, yet 69% of machine documents sit inside the human range. One project instruction file, loaded silently by the harness, moved a model from 6.10 dashes per 1,000 words to 0.00, a larger swing than any between models.

1. The question

An em dash appears in a paragraph and somebody calls the paragraph machine written. A list of 3 does the same. So does the word delve, the phrase not just X but Y and a run of short sentences. Every week this argument runs somewhere in public and every week it runs on the same missing number: how often each of those features shows up in prose that was written before any of these models existed.

The claim being made is a claim about a rate and a rate needs a denominator. This study measures both halves. It counts 10 such features in engineering prose published before 2020 and in prose written by 4 models on the same topics. Then it asks the second question that the public argument never reaches: whether a feature that really is more frequent in machine text is thereby able to classify a document.

Frequency and discriminating power are not the same quantity. A feature can run twice as often in one corpus and still overlap so heavily that no threshold sorts the documents. That gap is where the whole result sits.

2. Method

2.1 The threshold was fixed before the data existed

`CRITERIA.md` was committed before a single text was downloaded and it names what would count as a machine fingerprint:

> *a rate at least 10 times higher in corpus B **and** an AUC of 0.90 or better on the per document distributions.*

AUC here is the probability that a randomly chosen machine document carries a higher rate of the feature than a randomly chosen human document. 0.5 means the feature does not separate the 2 corpora at all. 1.0 means it separates them perfectly. Both halves of the threshold were chosen in advance and neither was touched afterwards. Corrections to the criteria are appended at the bottom of that file with dates and reasons rather than edited into the text above them.

2.2 Corpus A, human prose from before 2020

A document entered corpus A only if its publication date was verifiable from the URL or from a stamp in the body of the text and only if that date fell before 2020-01-01. Dates from HTML metadata were refused, because content management systems rewrite them on migration. The other gates: English, engineering prose rather than code or changelogs,

400 to 6,000 words after cleaning, no more than 25% of the text in code and a living author writing in public without a login.

Sources were read in the order their own index shows them, no more than 8 documents per author, so that nothing was picked for sounding a particular way. The result is Paul Graham, Joel Spolsky, Brendan Gregg, Julia Evans and 8 RFCs taken by descending number from 8690 downwards. RFCs are reported as a separate stratum, because a specification is not an essay and averaging them together would hide both.

	DOCUMENTS	WORDS
A, essays before 2020	32	48,914
A, RFCs before 2020	8	23,168

The corpus was assembled on 21 August 2026. 152 candidate documents were rejected by the gate. Every rejection carries its reason in `corpus-a.json`: 117 for a date at or after the cutoff, 25 for being under 400 words, 5 for exceeding 6,000 and 5 for carrying too much code. The texts themselves are not redistributed. What ships is the URL, the date, the word count and the sha256 of the cleaned text, with the download script beside it.

The 2 strata stay apart in every statistic below and the essays are the denominator. Each AUC, each bootstrap interval, each ratio and each mention of the human range compares the 32 essays against the machine documents. The 8 RFCs appear in the rate columns and nowhere else. Pooling them in would flatter the result rather than measure it: a specification runs at 0.04 dashes per 1,000 words against 1.33 for the essays, so adding it to the human side would widen every gap for free.

2.3 Corpus B, the same topics written by 4 models

Corpus B is the titles of the corpus A essays handed back to a model with a neutral instruction: write a blog post for a software engineering audience on this title, about 1,150 words, as prose in paragraphs. No style guidance in either direction, because asking for human sounding prose or for machine sounding prose would answer the question inside the prompt.

4 models, 30 texts each, the same 30 topics throughout, generated on 21 August 2026 with claude CLI 2.1.238, tools isolated with `--tools ""` and an empty MCP config.

2 of the sonnet texts, `B-002` and `B-021`, were generated twice: the first run stopped on a tool use stub of 13 and 17 words. `stats.json` records both as `failures` and the stubs sit in `rejected/`. The 2 replacements are the ones measured. Sonnet is the model with the weakest dash separation, so the reader should know the cell was retried.

The 30 topics are the first 30 corpus A essays in corpus order. The run was fixed at 30 texts per model before it started, so the 2 essays that sit last in that order, `A-031` and `A-032`, both by Julia Evans, have no machine counterpart. Nothing about those 2 texts caused the omission and both stay in corpus A.

	DOCUMENTS	WORDS
B, `claude-fable-5`	30	35,077
B, `claude-haiku-4-5-20251001`	30	33,182
B, `claude-opus-5`	30	35,704
B, `claude-sonnet-5`	30	33,364
B, all 4 pooled	120	137,327

2.4 The counter

`tells.mjs` holds 10 rules, copied rather than imported from the writing filter they came from, so that the version of the rules that produced these numbers stays frozen beside them. Rules tuned to one author's personal punctuation were dropped on purpose, because they encode a house preference rather than a claim about who wrote the text. One rule was added: `tricolon`, a coordinated list of exactly 3 short items, because the hypothesis names the rule of 3 explicitly and the source filter has no such rule.

Word counts use one definition across every document and every corpus, since the denominator decides the comparison. Every rule is a counter here rather than a verdict, which is the one change of role from the filter it came from.

3. Results

No tell of the 10 cleared the threshold. The best of them, `tricolon`, reached AUC 0.80 with a 95% bootstrap interval of 0.70 to 0.88 and a ratio of 2.6. The bar sits outside that interval, so this is not a near miss that more documents would rescue.

TELL	A: ESSAYS	A: RFCS	B: ALL 4 MODELS	RATIO	AUC	AUC 95% CI
`dash`	1.33	0.04	4.70	3.5	0.69	0.61 to 0.77
`tricolon`	1.35	1.12	3.54	**2.6**	**0.80**	0.70 to 0.88
`not-just-x-but-y`	0.22	0.04	0.11	0.5	0.42	0.35 to 0.49
`it-s-not-x-it-s-y`	0.02	0.00	0.04	1.8	0.51	0.46 to 0.54
`llm-vocabulary`	0.00	0.00	0.09	inf	0.55	0.52 to 0.57
`llm-vocabulary-figurative`	0.02	0.00	0.04	2.1	0.51	0.47 to 0.54
`hollow-opener`	0.02	0.00	0.01	0.4	0.49	0.45 to 0.51
`engagement-bait`	0.00	0.00	0.00	n/a	0.50	0.50 to 0.50
`hype-emoji`	0.00	0.00	0.00	n/a	0.50	0.50 to 0.50
`staccato-run`	0.00	0.00	0.00	n/a	0.50	0.50 to 0.50

Rates are matches per 1,000 words. Ratio is machine over human essays. The AUC column compares the 32 essays against the 120 machine documents.

The headline says the best of 10 and the honest count of testable rules is 7. 3 of the 10 rules produced 0 matches in all 160 documents: `engagement-bait`, `hype-emoji` and `staccato-run`. Their AUC of 0.50 is arithmetic rather than evidence, since 2 samples of identical zeros can only tie. In engineering prose the hype emoji and the comment bait ending are simply absent from both sides, so they carry no weak signal to argue about. That leaves 7 rules the data can rank and the best of those 7 is what fails the threshold. Saying so makes the result narrower and firmer: the 3 empty rules are not padding the failure. `llm-vocabulary` is the marginal case, 13 matches in corpus B against 0 in corpus A, which is why its ratio is infinite and its AUC is still 0.55.

3.1 The em dash, measured

All 4 models use the em dash more often than these human authors do, so the folklore has the direction right. It has the strength wrong.

	RATE PER 1,000	RATIO TO HUMANS	AUC	DOCUMENTS WITH NO DASH
humans, essays before 2020	1.33			34%
`claude-fable-5`	6.10	4.6	0.70	37%
`claude-opus-5`	4.99	3.8	0.81	10%
`claude-haiku-4-5-20251001`	4.01	3.0	0.67	27%
`claude-sonnet-5`	3.60	2.7	0.56	47%
all 4 pooled	4.70	3.5	0.69	30%

47% of `claude-sonnet-5` essays contain no dash at all, against 34% of the human ones, so the absence of a dash carries no information whatever. 69% of machine documents

sit inside the range spanned by the human documents, so the presence of one carries very little.

The reverse figure is the worse one and it is 100%. Every one of the 32 human essays sits inside the range spanned by the machine documents, on the dash and on `tricolon` alike. Every rate an essay by Graham, Spolsky, Gregg or Evans reaches falls between the lowest and the highest machine rate, so a threshold built on either tell cannot exclude a human document at all.

645 dashes were counted in corpus B and 643 of them are em dashes. The other 2 are en dashes inside the date range October 29-31, 1 in `claude-fable-5` `text-B-023` and 1 in `claude-opus-5` `text-B-023`. The rule counts the en dash on purpose. No spaced hyphen standing in for a dash was found anywhere in the corpus.

3.2 Why 0.80 is still a failure

1.35 per 1,000 words for humans against 3.54 for the machines. It is the most consistent tell across the 4 models, which land between 3.08 and 4.07. It is the only rule here whose per document distributions pull apart at all.

It fails on both halves of the threshold. 2.6 is not 10 and 78% of machine documents fall inside the human range, so no threshold on this feature sorts a single document reliably.

3.3 Do the models differ from each other more than machine differs from human?

This question was added to the criteria when the study was widened from 1 model to 4, before the new numbers were counted, with the answer to be reported either way. Separation is the distance of AUC from 0.5.

TELL	HUMAN AGAINST MACHINE	WIDEST MODEL PAIR	WHICH PAIR
`tricolon`	**0.30**	0.11	haiku against opus
`dash`	**0.19**	0.14	fable against sonnet
`not-just-x-but-y`	0.08	**0.12**	fable against haiku
`llm-vocabulary`	0.05	**0.07**	fable against sonnet
`llm-vocabulary-figurative`	0.01	**0.07**	haiku against opus
`it-s-not-x-it-s-y`	0.01	**0.05**	opus against sonnet
`hollow-opener`	0.01	0.02	fable against haiku

No. On the 2 tells that carry any signal, the human against machine gap is the wider one. The model spread is wider on the other 5, but those are rules whose rates round to zero in both corpora, so that comparison is between 2 kinds of noise.

4. The room the text was written in

The largest effect in this study belongs to neither the model nor the prompt.

The first version of this measurement used 1 model and reported a headline finding: not one em dash in 30 machine texts, against 1.33 per 1,000 words for humans, AUC 0.17. The em dash pointed the opposite way to the folklore and it was by far the most quotable number here.

It was an artefact. Those 30 texts were generated from a working directory inside a private repository whose `CLAUDE.md` bans the em dash in English outright. The claude CLI loads

that file into the model's context by itself. The prompt asked for nothing about style. The room did.

`CLAUDE-FABLE-5`, SAME PROMPT, SAME 30 TOPICS	DASHES PER 1,000 WORDS	DOCUMENTS WITH NONE
generated with that `CLAUDE.md` loaded	0.00	100%
generated in a directory with no project instructions	**6.10**	37%

That swing, 6.10 against 0.00, is larger than any gap between 2 models in this study and larger than the gap between human and machine. **On these features, what is in the working directory matters more than which model is running.**

The contaminated corpus is kept in the repository rather than deleted and `context-check.mjs` is the check that catches this class of mistake: it walks from the working directory upward listing every file the CLI picks up as context and it records what it found for the clean directory and the dirty one alike. A check that has never returned a hit is not evidence of anything.

The second consequence lands on the writing filter the rules came from. Such a filter catches deviation from a house voice. It does not catch machine authorship and nothing measured here supports calling it a detector of anything else.

5. Threats to validity

Corpus A is biased toward survivors. Graham, Spolsky, Gregg and Evans are in it because their writing is still online and still dated, which is close to the definition of a text that lasted. The measurement describes engineering prose that people still read, not written English in general.

Corpus B comes from 1 prompt and 1 length target. Another instruction, say a request for a post aimed at a professional network, would very likely produce different rates.

The rules are regular expressions and they approximate what they are named after. `tricolon` finds a punctuation shape, not a rhetorical figure: it misses 3 parallel sentences and 3 parallel clauses without commas. `staccato-run` counts characters between full stops. False positives exist in both corpora and the assumption that they land evenly on both sides is untested here.

32 human essays against 120 machine documents is a small study, which is why the bootstrap interval is reported next to every AUC rather than the point estimate alone. The interval was not part of the registration, which fixed a threshold rather than a significance test.

Model identity is verified through the CLI's usage report and that verification is weaker for `claude-haiku-4-5-20251001` than for the other 3, because the CLI runs the same model for its own internal calls and its key is therefore always present.

The context check is a filesystem scan and it reported the generation directory clean. An advisory self report was collected alongside it, in which 1 of the 4 models answered that an empty auto memory index was loaded. The scan found no file carrying instructions, so nothing in the corpus is attributed to that, but the disagreement is on the record rather than smoothed over.

6. Prior work

The vocabulary half of this question has a strong published answer and this study is not first to it.

- Kobak, González-Márquez, Horvát and Lause, *Delving into LLM assisted writing in biomedical publications through excess vocabulary* (arXiv:2406.07016, Science Advances 11(27), 2025, doi:10.1126/sciadv.adt3813) is the closest neighbour and the most rigorous work in the area. It tracks word frequencies across more than 15 million PubMed abstracts from 2010 to 2024 and estimates that at least 13.5% of 2024 abstracts were processed with a language model. It measures words, in one corpus, over time. It does not measure punctuation or syntactic constructions and it does not report how well any single word classifies a document.

- Liang, Yuksekgonul, Mao, Wu and Zou, *GPT detectors are biased against non-native English writers* (arXiv:2304.02819, Patterns, 2023) is the standing result on what such features cost when they are wrong: several widely used detectors consistently label writing by authors whose first language is not English as machine generated. It evaluates deployed detectors rather than the individual features people argue about.
- *AI generated text detection: a comprehensive review of methods, datasets and applications* (doi:10.1016/j.cosrev.2025.100793, Computer Science Review, 2025) surveys the detector literature. The field is built around trained classifiers over full documents, which is a different object from a per feature base rate.
- Public detection datasets on Hugging Face, such as `artem9k/ai-text-detection-pile`, are labelled human and machine pairs for training classifiers. They are not frequency tables and they carry no date gate that would separate prose written before 2020.

What was not found is a frozen, dated corpus of prose written before language models existed, with per feature frequencies and an overlap statistic reported next to each one. The public argument about whether the em dash marks machine writing has run for over a year without a shared number to argue over. Searched arXiv, Semantic Scholar and GitHub on 29 August 2026, along with Hugging Face and Zenodo on the same day. Semantic Scholar rate limited most queries and general web search was unavailable that day, so the open web outside those sources is not claimed as checked. The general web search that produced the list above was run on 27 August 2026, which is 6 days after the measurement rather than before it.

7. Availability

Corpus B in full, the corpus A manifest with dates and hashes, the rule file, the analysis script, the machine readable statistics and the contaminated corpus kept as an exhibit are in the repository. The archived release carries the DOI above.

8. References

1. Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, Jan Lause. Delving into LLM assisted writing in biomedical publications through excess vocabulary. Science Advances 11(27), 2025. doi:10.1126/sciadv.adt3813. arXiv:2406.07016. https://doi.org/10.1126/sciadv.adt3813
2. Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, James Zou. GPT detectors are biased against non-native English writers. Patterns 4(7), 100779, 2023. doi:10.1016/j.patter.2023.100779. arXiv:2304.02819. https://doi.org/10.1016/j.patter.2023.100779
3. Tanzila Kehkashan, Raja Adil Riaz, Ahmad Sami Al-Shamayleh, Adnan Akhunzada, Noman Ali, Muhammad Hamza, Faheem Akbar. AI generated text detection: a comprehensive review of methods, datasets and applications. Computer Science Review 58, 2025. doi:10.1016/j.cosrev.2025.100793. https://doi.org/10.1016/j.cosrev.2025.100793
4. artem9k. `ai-text-detection-pile`. Hugging Face dataset, 2023. https://huggingface.co/datasets/artem9k/ai-text-detection-pile
5. Dmitriy Semenkevich. `tells-baseline`: corpus, preregistration, rule file and raw counts. GitHub, 2026. https://github.com/dimhold/tells-baseline

---

CITE THIS

```
Semenkevich, D. (2026). No writing tell separates human from machine on its own: the best of 10 reaches AUC 0.80. dimhold.by. https://doi.org/10.5281/zenodo.22128847
```

machine generated text detection · stylometry · corpus linguistics · AI writing tells · em dash · pre-2020 baseline · LLM evaluation