

What Makes Writing Sound AI Generated? A Study of Public Accusations, Writing Patterns and False Alarms

We read 787 public accusations that a text was written by AI, compared the accused texts with unchallenged ones, and measured how often a checker raises a flag on ordinary writing. No single cue is proof.

MAIN FINDING

Most accusers gave no reason, and no single cue is proof. The em dash character was far more common in challenged short comments than in unchallenged ones, and showed no difference in long articles. Every pattern we could measure in older writing also appears in comments from before ChatGPT.

- Observational study complete: 787 accusations, 967 unchallenged comparison texts
- Checker false-alarm test complete
- Reading experiment designed, not run
- Non-native English writers not tested

PUBLISHED	COLLECTED	VERSION
October 12, 2026	Hacker News, January 2023 to October 2026	1.1 (updated October 13, 2026)

EVIDENCE

Public posts, text pattern counts, blind edit ratings

Canonical URL: <https://ultratextgen.com/research/ai-writing-footprint-study/>
UltraTextGen Research · Version 1.1 · Published October 12, 2026 · Updated October 13, 2026

SUMMARY

What the study found

No single writing cue is proof that a text was written by AI. People do accuse text of being AI-written in public, but they rarely say why, and the patterns we could measure in

older writing also turn up in comments from before ChatGPT. We did not measure how often AI text has these patterns, so this study cannot say how good any cue is, only that none is proof.

57%

of 787 accusations name no reason at all

451 accusers gave none by the original coding (437 and 518 when two further blind codings were run, so between 55% and 66%). The most common reasons given were that the text says very little (74 accusers), a feeling about its tone (68) and that it is off-topic or wrong (42). Only 22 of 287 comment accusers (7.7%) and 66 of 426 article accusers (15.5%) named a feature of the wording or formatting. The em dash was named 18 to 19 times, or about 2.4%, by every coding.

● COUNTED IN PUBLIC POSTS

◇ MODEL-ASSISTED LABELS

18% vs 2%

of challenged short comments contain the em dash character, against unchallenged ones

52 of 287 challenged comments against 12 of 564 matched unchallenged ones (odds ratio 8.4 across the 229 matched sets). In long articles it is 47% against 46%. In ordinary comments from 2019 to 2021, before ChatGPT, 1.2% had one. The typed double hyphen (--) shows no link at all.

● COUNTED IN PUBLIC POSTS

52%

of unchallenged articles got a flag from our checker

At 1,200 words or more it was 64%. On 300 stretches of books published before 1910 it was 21%. For challenged articles it was 65%. Once the kind of article is held fixed the gap between the two shrinks from an odds ratio of 1.7 to 1.25 and its interval includes 1.

● TOOL RUN ON REAL TEXT

1 of 81

passages read better after the checker's dash replacement

Two blind raters compared 81 passages with every flagged dash replaced against the original. The edit won once for each rater. In the 36 modern passages the original won 27 and 23 times.

◇ AI-MODEL RATERS

○ NOT TESTED What this study cannot tell you

It cannot tell you who wrote any text. No text here has independently verified authorship, and the study does not try to detect AI. It cannot show that any pattern causes suspicion: that needs a reading experiment, which we designed and did not run. It does not cover Reddit or any platform other than Hacker News, Lemmy, Discourse and Stack Exchange, and it did not test writers who use English as an additional language. UltraTextGen publishes this study and makes the checker it tests. Read the limitations

YOUR WRITING

What this means for you

One em dash is not a verdict. Ordinary writers use it, and replacing every flagged dash with a comma, colon or full stop did not make the passages we tested read better. Whether removing other patterns helps was not tested.

Specific beats smooth. The reason accusers gave most often was that the text says very little. Name one thing a reader could check: a number, a name, a date, a place. That helps whoever wrote it.

Short sentences in a row stand out most. Three or more sentences of six words or fewer, one after another, was the most consistent pattern. Use them where they fit and mix in a longer one where they do not.

Cleaning patterns will not prevent an accusation. 27% of challenged comments carried none of the 21 patterns we measured, and more than half of accusers named no reason. If you are accused, expect an argument about tone, not proof: in the threads we read, 31% of accusations on comments drew a reply disagreeing with the accuser, and the accused author denied it in 16%. Those are replies, not verdicts.

Do not copy patterns to look human. Cutting a pattern to avoid a flag is a different goal from writing well. Edit for the reader.

The AI Writing Footprint Checker shows you which wording patterns a reader might notice. It runs in your browser, sends your text nowhere and gives no score or probability of AI authorship. This study found parts of it that need work, listed under checker false alarms.

DESIGN

Questions and design

We asked which writing patterns make readers call a text AI-written, how often those patterns appear in ordinary writing, and how many false alarms a pattern checker gives.

Three things are kept apart throughout. **Pattern detection** means a measurable feature exists in the text.

Reader perception means someone said the text sounds like AI. **Actual authorship** means how the text

was made. This study measures the first two. It never measures the third.

QUESTION	ANSWERED?	WHERE
WHAT MAKES WRITING APPEAR AI-GENERATED?	What accusers say: yes. What causes it: no	What accusers say
WHICH PATTERNS DO ACCUSERS NAME MOST?	Yes, with counts and intervals	Figure 1
WHICH CUES CREATE THE STRONGEST PERCEPTION?	Not tested. Association only	What we did not test
DO COMBINATIONS MATTER MORE?	They add up; none is special. Association only	Do cues add up?
HOW COMMON ARE THE PATTERNS IN ORDINARY WRITING?	Yes for comments by period, books before 1910 and articles by kind. Not for non-native writers	Ordinary human writing
HOW MANY FALSE ALARMS DOES THE CHECKER GIVE?	Yes, by rule, length, kind and period	Checker false alarms
WHAT SHOULD WRITERS CHANGE?	Partly: edits judged by blind raters	Does editing help?

Design in short

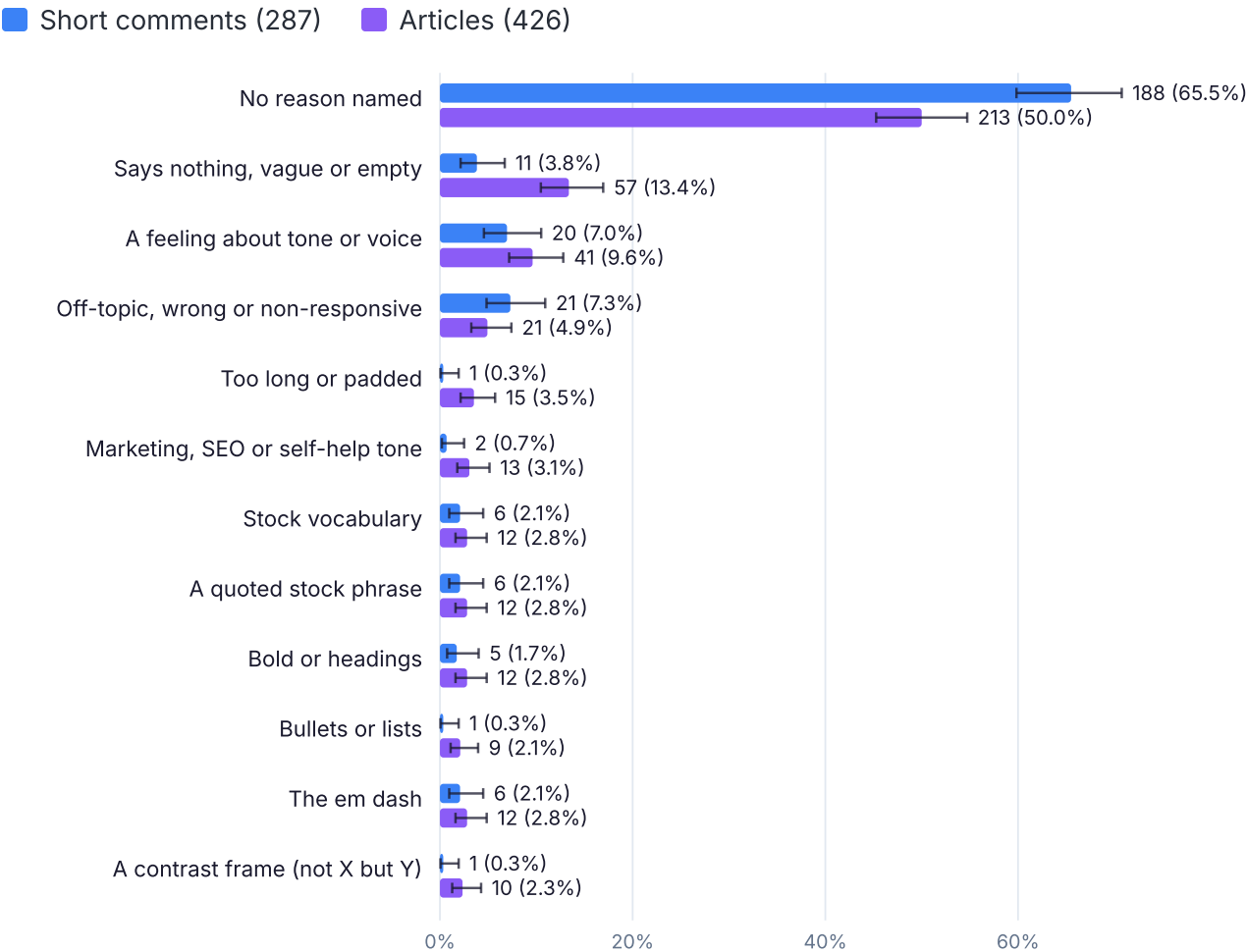
We searched Hacker News from January 2023 to October 2026 for replies that accuse a specific comment or article of being AI-written, kept the cases where the accuser meant a specific text, and recorded what the accuser named. We also kept 74 cases from Lemmy, Discourse and Stack Exchange, which we describe and never compare. For each accused comment we picked up to three unchallenged comments from the same thread with similar length. For each accused article we picked up to two unchallenged front-page articles from the same days with similar popularity and length. We then counted the same patterns in both groups and compared them. “Unchallenged” means nobody accused the text. It does not mean human.

RESULTS

What accusers say

Most accusers name no reason. Of those who do, the commonest is that the text says very little.

Figure 1. What accusers say. Share of accusers who named each reason, with 95% intervals.



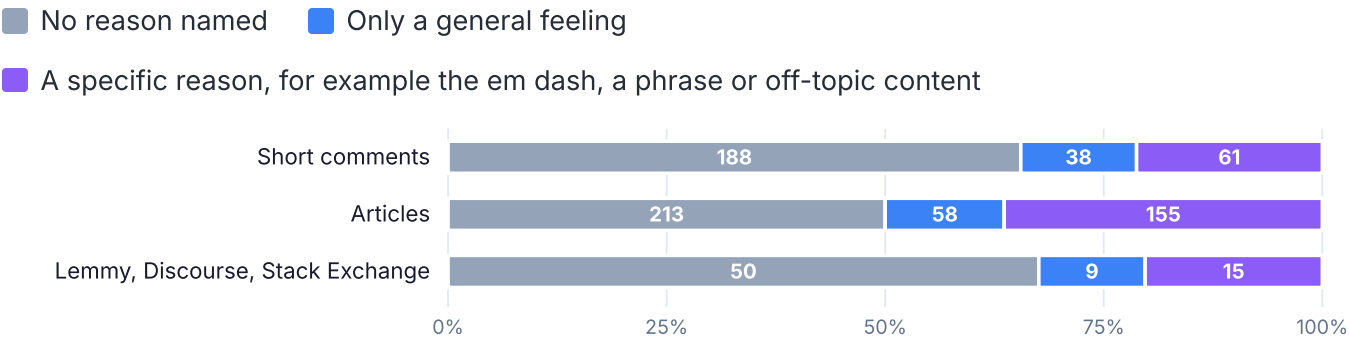
An accuser can name more than one reason, so reasons do not add to 100%. Counts depend on the search wording we used.

▼ Data behind this chart

REASON	COMMENT ACCUSERS	ARTICLE ACCUSERS
NO REASON NAMED	188 of 287	213 of 426
SAYS NOTHING, VAGUE OR EMPTY	11 of 287	57 of 426
A FEELING ABOUT TONE OR VOICE	20 of 287	41 of 426
OFF-TOPIC, WRONG OR NON-RESPONSIVE	21 of 287	21 of 426
TOO LONG OR PADDED	1 of 287	15 of 426
MARKETING, SEO OR SELF-HELP TONE	2 of 287	13 of 426

REASON	COMMENT ACCUSERS	ARTICLE ACCUSERS
STOCK VOCABULARY	6 of 287	12 of 426
A QUOTED STOCK PHRASE	6 of 287	12 of 426
BOLD OR HEADINGS	5 of 287	12 of 426
BULLETS OR LISTS	1 of 287	9 of 426
THE EM DASH	6 of 287	12 of 426
A CONTRAST FRAME (NOT X BUT Y)	1 of 287	10 of 426

Figure 2. Most accusations name nothing specific. Accusations by what the accuser named.



▼ Data behind this chart

GROUP	NO REASON NAMED	ONLY A GENERAL FEELING	A SPECIFIC REASON
SHORT COMMENTS	188	38	61
ARTICLES	213	58	155
LEMMY, DISCOURSE, STACK EXCHANGE	50	9	15

Short comments: 188 of 287 accusers named nothing, and only 22 (7.7%) named a feature of the wording or formatting (33, or 11.5%, if "says nothing" counts as one). Articles: 213 of 426 named nothing, and 66

(15.5%) named such a feature (109, or 25.6%, with “says nothing”). The off-topic and tone reasons are judgments about content and feeling, not wording, and no pattern rule can read them.

Of the 426 article accusations, 28 are posts submitted to Hacker News as a link, where the accuser was reacting to the linked page and we analysed only the short post, and 21 are text posts. Those 49 count here, in what accusers said, and in no statistic about the text itself.

Accusations are argued with. 89 of 287 comment accusations (31%) drew at least one reply that disagreed, and 41 (14%) one that agreed. In 47 cases (16%) the author of the accused comment replied to deny using AI. In 10 (3%) the author said they used a tool for proofreading, polish or translation, and in 3 (1%) that an AI wrote it. These are self-reports we could not verify. A reply is not a vote: most agreement on Hacker News is an upvote nobody sees.

Many challenged texts show none of the patterns. 78 of 287 challenged comments (27%) carried none of the 21 pattern flags in our table, against 269 of 564 unchallenged comments (48%). Most challenged comments had at least one, but they also had accusers who named something else, or nothing.

The counts depend on how we searched. 8 of our first 52 search phrases named cues such as “em dash”, our filter cannot see accusations worded without an AI term, and from February 2026 the monthly searches hit the search service’s cap of 1,000 results. The labels are model-made. An independent re-check of 60 random cases found at least one coding error in 10 of them (16.7%, interval 9.3% to 28.0%), 4 of them cases that should not have counted as an accusation of the shown text. Treat every count here as approximate. We cannot say how often readers accuse a given text, only what was said when they did.

RESULTS

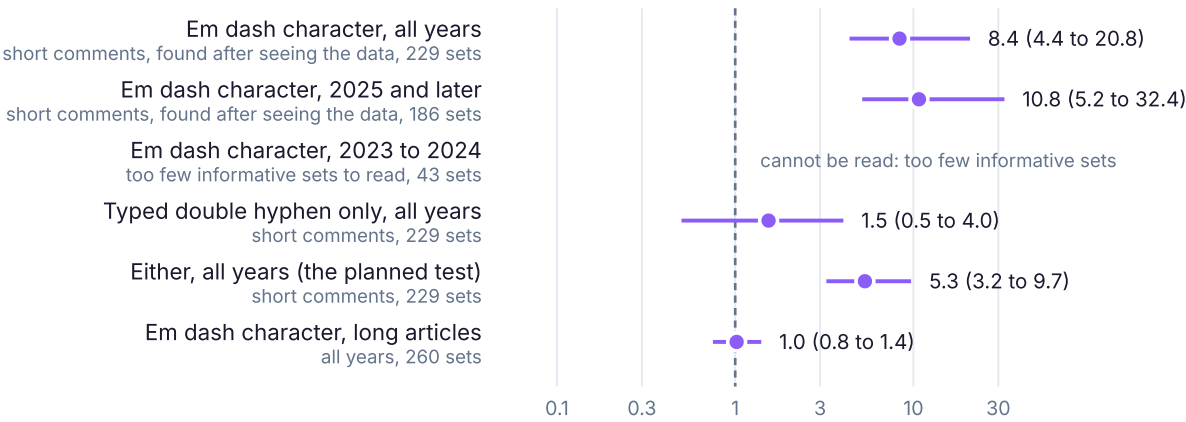
The em dash

The em dash is the most talked-about cue. The character itself is far more common in challenged short comments, but it is not a rule: it is absent from most challenged comments, and it makes no difference in long articles.

How to read an odds ratio. 1 means no difference between challenged and unchallenged texts. 5 means the pattern had about five times the odds of appearing in a challenged text. It describes how often the pattern shows up. It does not say what caused anyone to accuse.

Two different marks. Our planned flag counted the em dash character and a typed double hyphen between words (--). They behave differently, so Figure 3 splits them. The split was made after we saw the data, which makes the character result a lead to replicate, not a settled finding.

Figure 3. The em dash and being accused. Matched odds ratios with 95% intervals, log scale. 1 means no difference.



Intervals are percentile bootstrap intervals over matched sets. They are unreliable with few informative sets: for 2023 to 2024 there are 4 for the character and 8 for either mark, and a different method gives a much wider interval. The 2023 to 2024 row means we cannot say, not that there is no link. The split between the character and the double hyphen was made after seeing the data.

▼ Data behind this chart

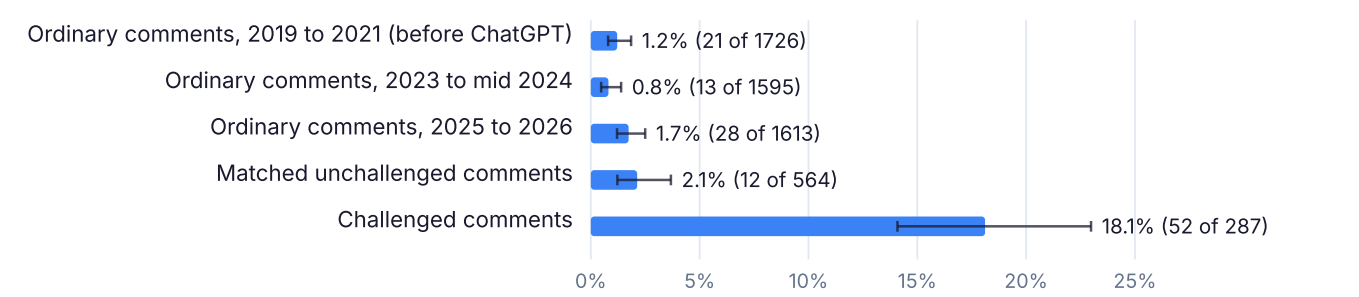
GROUP	MATCHED SETS	ODDS RATIO	95% INTERVAL	CHALLENGED WITH IT	UNCHALLENGED WITH IT
EM DASH CHARACTER, ALL YEARS	229	8.37	4.39 to 20.8	41 of 229	12 of 564
EM DASH CHARACTER, 2025 AND LATER	186	10.76	5.17 to 32.42	40 of 186	9 of 457
EM DASH CHARACTER, 2023 TO 2024	43	0.67	not estimable	1 of 43	3 of 107
TYPED DOUBLE HYPHEN ONLY, ALL YEARS	229	1.54	0.5 to 4.05	7 of 229	12 of 564
EITHER, ALL YEARS	229	5.35	3.25 to 9.72	48 of 229	24 of 564

GROUP	MATCHED SETS	ODDS RATIO	95% INTERVAL	CHALLENGED WITH IT	UNCHALLENGED WITH IT
EM DASH CHARACTER, LONG ARTICLES	260	1.02	0.75 to 1.4	122 of 260	186 of 403

Short comments. 52 of 287 challenged comments (18.1%) contain the em dash character, against 12 of 564 matched unchallenged ones (2.1%). The matched odds ratio is 8.4 over all years and 10.8 from 2025. The typed double hyphen shows nothing (1.5, interval 0.5 to 4.1). **2023 and 2024:** the data cannot show a link either way. Only 4 matched sets were informative for the character and 8 for either mark, so the interval is unreliable and the honest reading is “unknown”, not “absent”. **Long articles:** 47% of challenged and 46% of unchallenged articles have the character (odds ratio 1.0). Short comments and long articles differ in many ways at once, among them length, audience and how people write them, so we cannot say which one explains the difference.

An educated guess, labeled as one: in short comments the dash may have become a quick, nameable reason once it was widely called a tell. The data fit that, but they also fit the possibility that comments challenged in 2025 and later were more often machine-written. Nothing here separates the two.

Figure 4. How often comments contain the em dash character. Share of comments with at least one, with 95% intervals.



The three ordinary-comment rows keep comments that mention AI; dropping them changes no share by more than 0.2 points. Matched unchallenged comments never mention AI by design. Comments from 2023 on are not known to be human-written; only the 2019 to 2021 row is certain. Typed double hyphens fell from 2.4% to 1.2% over the same periods, which is why the combined figure looks like a fall.

▼ Data behind this chart

GROUP	COMMENTS	WITH THE CHARACTER	SHARE	95% INTERVAL
ORDINARY COMMENTS, 2019 TO 2021 (BEFORE CHATGPT)	1726	21	1.2%	0.8 to 1.9
ORDINARY COMMENTS, 2023 TO MID 2024	1595	13	0.8%	0.5 to 1.4
ORDINARY COMMENTS, 2025 TO 2026	1613	28	1.7%	1.2 to 2.5
MATCHED UNCHALLENGED COMMENTS	564	12	2.1%	1.2 to 3.7
CHALLENGED COMMENTS	287	52	18.1%	14.1 to 23.0

What accusers say matches the gap between talk and data. They named the dash in 0 of 151 accusations in 2023 and 2024, 9 of 240 in 2025 and 10 of 395 in 2026. With so few early cases, a rate as high as 2.5% in 2023 and 2024 would not be ruled out, so we do not claim the talk began in 2025. In short comments, 82% of challenged comments had no em dash character.

RESULTS

Which patterns held up

One pattern, runs of short sentences, was consistent across comments and articles. Everything else was weak, time-bound, or went the other way.

Supported means the link held in both comments and articles after the checks described under method. Where the rules of those checks and our own more cautious reading disagreed, we used the more cautious one. **Weakly supported** means it held in one place, in a single period, or only before an adjustment. **Contradicted** means the data point the other way or show no link where one is often claimed. **Untested** means the data cannot say. Odds ratios are for comments, then articles.

PATTERN	STATUS	WHAT THE DATA SHOW
RUNS OF THREE OR MORE SHORT SENTENCES	SUPPORTED	Short comments 6.4; articles 2.7. Holds with genre held fixed (2.3) and with same-genre matched controls only (2.7, interval 1.3 to 7.4). Present in about 1 in 60 to 90 ordinary comments. Short sentences are also a common style in dialogue and chat; we did not test that.
THE EM DASH CHARACTER	WEAKLY SUPPORTED	Short comments: 8.4 over all years and 10.8 from 2025 (a split we made after seeing the data). The data cannot show a link in 2023 to 2024. Long articles: 1.0. The character is in 1 to 2% of ordinary comments in every period.
TYPED DOUBLE HYPHEN (--) USED AS A DASH	CONTRADICTED	No association in short comments (1.5, interval 0.5 to 4.1) and none in articles (1.0).
LISTS OF THREE ITEMS	WEAKLY SUPPORTED	Short comments 2.2; inconclusive in articles. Our test matches any three-item list, including ordinary ones.
QUESTION FOLLOWED BY A SHORT ANSWER	WEAKLY SUPPORTED	Short comments 4.0; articles 1.8, falling to 1.3 once genre is held fixed. The test also matches ordinary dialogue.

PATTERN	STATUS	WHAT THE DATA SHOW
BULLET LISTS	◇ WEAKLY SUPPORTED	Articles 2.1 (1.9 with genre held fixed; 1.7 with same-genre controls, interval 0.6 to 5.0). Normal in documentation. Bold and headings could not be measured, because Hacker News does not show them.
"HERE'S THE THING" SCAFFOLDING	◇ WEAKLY SUPPORTED	Articles 2.4, a pattern we looked for after seeing the data, so a lead to replicate.
REPEATED SENTENCE OPENINGS	◇ WEAKLY SUPPORTED	Articles 1.9, falling to about 1.2 to 1.4 once page structure or genre is held fixed.
GENERIC OPENING PRAISE	◇ WEAKLY SUPPORTED	8 challenged comments against 1 unchallenged. Too few to rank.
"NOT ONLY X BUT Y" AND "IT'S NOT X, IT'S Y"	▣ CONTRADICTED IN ARTICLES	Strict forms: articles 1.2. Short comments 3.4, which did not survive adjustment. These are two different constructions, and our test groups them.
FORMAL TRANSITIONS (MOREOVER, FURTHERMORE)	▣ CONTRADICTED	Challenged articles had fewer of them (0.5).
SEMICOLONS	▣ CONTRADICTED	No association (1.1).
ELLIPSES	▣ CONTRADICTED	Challenged comments had fewer (0.5).
STOCK PHRASES ("PLAYS A CRUCIAL ROLE")	○ UNTESTED	Too rare in both groups to estimate (3 against 1 in comments).
UNIFORM SENTENCE LENGTH	○ UNTESTED AS A CAUSE	A rule for it fires more on challenged comments (22% against 9%) but also on 21% of comments written before ChatGPT. One accuser in 787 named it.
VAGUE STORYTELLING, POLISHED WORDING, UNNATURALLY EVEN TONE	○ UNTESTED	No deterministic test. These are what accusers mean by "says nothing" and "tone", and no rule can read them.

PATTERN	STATUS	WHAT THE DATA SHOW
BALANCED ARGUMENTS, GENERIC SUMMARIES, REPETITIVE PARAGRAPH SHAPE	UNTESTED	Too rare or not measurable in comments.

The full table with every pattern, count and interval is in `pattern-associations.csv`. We defined 30 patterns and list the 21 primary flags in the table; some were planned in advance (14) and some added after we saw the data (7), and the later ones are never ranked above “weakly supported”. The corrections for the number of tests (Benjamini-Hochberg) covered 29 flags per kind of text, variants included. The period, genre and checker-rule comparisons are not, so read them as leads.

RESULTS

Do cues add up?

Several weak cues together said more than any one cue, but no particular pair or trio was shown to be special.

In short comments, the number of different patterns in a text predicted being challenged beyond the em dash alone, and the em dash still added something on top. In the model we fitted, the effects added; they did not multiply (interaction $p = 0.79$). That is one model, and some single patterns did interact with the pattern count: a question followed by a short fragment ($p = 0.005$) and, in articles, the “X, not Y” construction ($p = 0.04$). In a repeated hold-out test, the em dash alone improved prediction of held-out comments by 12.7 log-likelihood units over a length-and-topic baseline, the pattern count alone by 17.0, and both by 20.0. In articles the pattern count mattered less (about 1.2 times the odds per extra pattern). Pairs were too sparse to rank: a question followed by a short fragment together with a run of short sentences occurred in 9 challenged comments and in none of 564 unchallenged ones, which is too few to say more.

An educated guess: readers may react to a general feel of templated writing instead of to any one mark. This is an association in a convenience sample, not an effect we measured on readers, and the pattern count also tracks how much text there is.

RESULTS

Ordinary human writing

Every pattern we tested appears in writing nobody challenged, and most were as common in comments from before ChatGPT as after.

We drew 4,934 ordinary Hacker News comments of 40 to 400 words at random from three periods, from random one-hour windows, and counted the same patterns. Only the first row is certain to be human-written; the other two periods are simply ordinary comments. Comments that mention AI are kept in the rows below. Leaving them out changes no share by more than 0.5 points; the amount left out grows with time (1% of eligible comments in 2019 to 2021, 6% in 2023 to 2024, 15% in 2025 to 2026).

PERIOD	EM DASH	3+ SHORT SENTENCES IN A ROW	"X, NOT Y"	QUESTION + SHORT FRAGMENT	BULLET LIST	COMMENTS
2019 TO 2021	3.5% (2.8 to 4.5)	1.6% (1.1 to 2.3)	4.2% (3.3 to 5.2)	2.8% (2.1 to 3.7)	3.9% (3.1 to 4.9)	1726
2023 TO MID 2024	2.4% (1.8 to 3.3)	1.4% (0.9 to 2.1)	3.4% (2.6 to 4.4)	2.6% (2.0 to 3.5)	2.1% (1.5 to 2.9)	1595
2025 TO 2026	2.9% (2.2 to 3.9)	1.2% (0.8 to 1.9)	4.8% (3.8 to 5.9)	2.4% (1.8 to 3.3)	3.2% (2.4 to 4.1)	1613

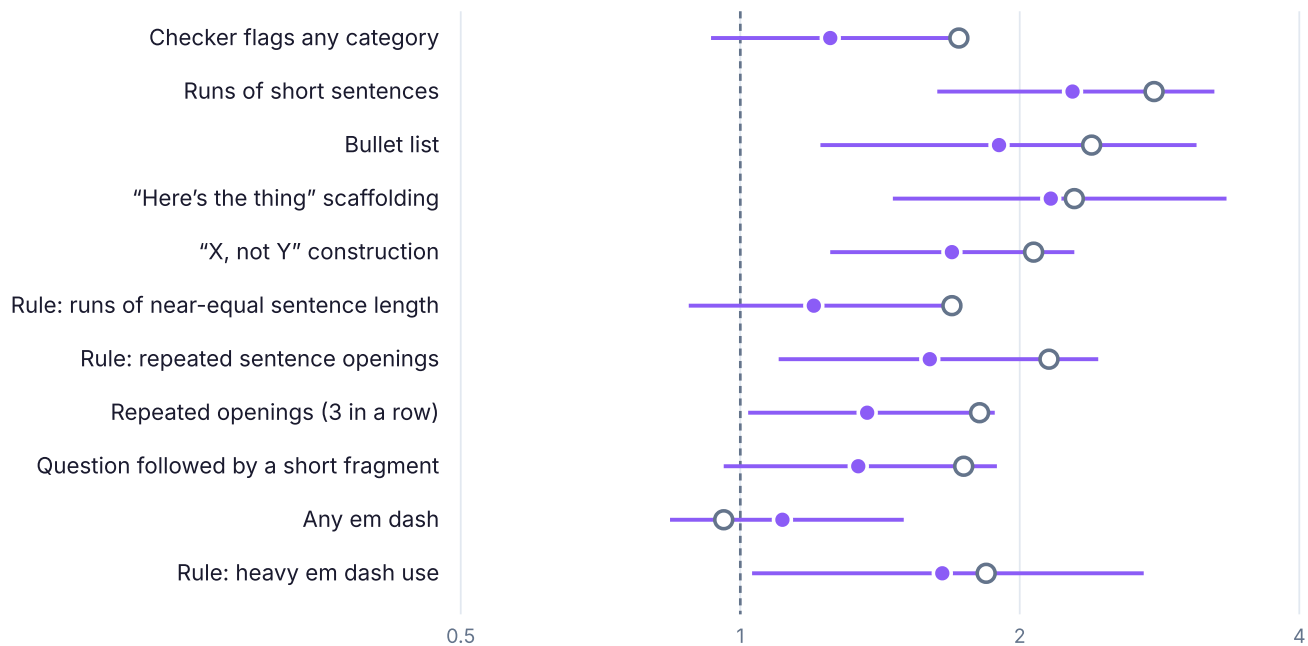
Percentages are shares of comments, with 95% intervals. "Em dash" here means the character or a typed double hyphen, the way our planned flag counted it. The character alone is 1.2%, 0.8% and 1.7% across the three periods (Figure 4); typed double hyphens fell from 2.4% to 1.2%, which accounts for the apparent fall in the first column.

Books published before 1910 are the cleanest human baseline there is. Ten books (Austen, Dickens, Doyle, Darwin, Shelley, Stoker, Twain, Carroll, Melville and a 1908 translation of Machiavelli) gave 300 chunks of 400 words; the checker flagged 63 of them. The heavy-dash rule was one of the rules raised in 33 of those 63 chunks (rules overlap). These are also the books the checker's thresholds were first tuned on, so the result is a floor on false alarms, not a fresh test. Dashes in these books follow period printing habits and a transcription, not only the authors' choices.

The kind of article matters more than we first allowed. Of the 377 challenged articles with a linked page, 45% are personal blogs (171) against 27% of the 403 unchallenged ones (108), and 5% technology press (20) against 18% (71). Our matching used popularity and length, not the kind of article. Comparing within kinds, some links shrink a lot and some hold:

Figure 5. Challenged articles versus unchallenged articles, before and after holding genre fixed. Odds ratios on a log scale.

○ Without genre ● Genre held fixed (interval shown)



This comparison uses the 377 challenged articles that have a linked page and all 403 unchallenged ones, and it ignores the matched sets, so its first column is not the matched odds ratio used elsewhere (1.8 for any category). It shows what genre alone does. Keeping the matched sets and using only controls of the same genre as the challenged article gives 2.1 for any category, with a very wide interval (0.75 to 7.1, 67 sets), and 2.7 (1.3 to 7.4) for runs of short sentences.

▼ Data behind this chart

MEASURE	WITHOUT GENRE	95% INTERVAL	GENRE HELD FIXED	95% INTERVAL
CHECKER FLAGS ANY CATEGORY	1.72	1.31 to 2.31	1.25	0.93 to 1.74
RUNS OF SHORT SENTENCES	2.79	2.03 to 3.88	2.28	1.63 to 3.24
BULLET LIST	2.39	1.54 to 3.80	1.90	1.22 to 3.10
"HERE'S THE THING" SCAFFOLDING	2.29	1.61 to 3.34	2.16	1.46 to 3.34
"X, NOT Y" CONSTRUCTION	2.07	1.57 to 2.78	1.69	1.25 to 2.29

MEASURE	WITHOUT GENRE	95% INTERVAL	GENRE HELD FIXED	95% INTERVAL
RULE: RUNS OF NEAR-EQUAL SENTENCE LENGTH	1.69	1.28 to 2.28	1.20	0.88 to 1.67
RULE: REPEATED SENTENCE OPENINGS	2.15	1.51 to 3.17	1.60	1.10 to 2.43
REPEATED OPENINGS (3 IN A ROW)	1.81	1.37 to 2.42	1.37	1.02 to 1.88
QUESTION FOLLOWED BY A SHORT FRAGMENT	1.74	1.30 to 2.42	1.34	0.96 to 1.89
ANY EM DASH	0.96	0.73 to 1.28	1.11	0.84 to 1.50
RULE: HEAVY EM DASH USE	1.84	1.21 to 2.89	1.65	1.03 to 2.72

The checker's overall flag went from an odds ratio of 1.7 to 1.25, with an interval (0.93 to 1.74) that now includes 1. With the matched sets kept and only controls of the same kind of article used, it is 2.1 with a very wide interval (0.75 to 7.1), so the honest reading is "weaker than the matched figure first suggested, and uncertain". Runs of short sentences fell from 2.8 to 2.3 and stayed above 1 in every version. Repeated openings and the question-plus-fragment pattern fell to about 1.4 and 1.3. The em dash stayed near 1. A model assigned the kind of article from the name of the website alone, without being told which articles were challenged; about half of its labels were marked low-confidence, and 108 articles fell in an unknown group. This is a check on robustness, not a replacement estimate.

Not covered: writers who use English as an additional language, professional marketing copy at scale (70 company-blog articles), academic prose at scale (15 articles) and personal essays outside blogs.

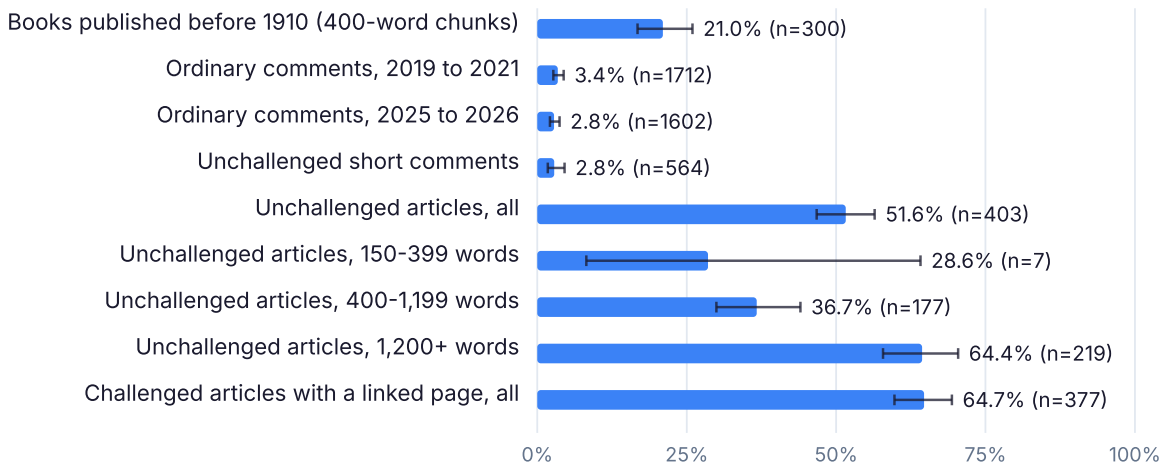
FALSE ALARMS

How many false alarms does the checker give?

On ordinary long articles and on old books, the checker raised a flag far more often than a reader should take as a warning. On short comments it almost never did.

We ran all 22 rules of the AI Writing Footprint Checker over every challenged and unchallenged text, over the books, and over the ordinary comments. A "flag" here means a category rated noticeable or high. The checker gives no score and says nothing about authorship; this shows how often its rules fire on writing of known or ordinary origin. Only the books and the 2019 to 2021 comments are known to predate AI. Every other row is writing nobody challenged, which is not the same as known to be human.

Figure 6. How often the checker raised at least one category (noticeable or high). By kind of text, with 95% intervals.



Books published before 1910 cannot be AI-written, so every flag on them is a false alarm (the interval treats each chunk as independent; counting each of the ten books as one unit gives 12% to 31%). These ten books are also the set the checker’s thresholds were first tuned on. Unchallenged texts are not known to be human-written. Challenged articles are the 377 with a linked page.

▼ Data behind this chart

TEXT	TEXTS	RAISED A CATEGORY	SHARE	95% INTERVAL
BOOKS PUBLISHED BEFORE 1910 (400-WORD CHUNKS)	300	63	21.0%	16.8 to 26.0
ORDINARY COMMENTS, 2019 TO 2021	1712	59	3.4%	2.7 to 4.4
ORDINARY COMMENTS, 2025 TO 2026	1602	45	2.8%	2.1 to 3.7
UNCHALLENGED SHORT COMMENTS	564	16	2.8%	1.8 to 4.6
UNCHALLENGED ARTICLES, ALL	403	208	51.6%	46.7 to 56.5
UNCHALLENGED ARTICLES, 150-399 WORDS	7	2	28.6%	8.2 to 64.1
UNCHALLENGED ARTICLES, 400-1,199 WORDS	177	65	36.7%	30.0 to 44.0
UNCHALLENGED ARTICLES, 1,200+ WORDS	219	141	64.4%	57.8 to 70.4

TEXT	TEXTS	RAISED A CATEGORY	SHARE	95% INTERVAL
CHALLENGED ARTICLES WITH A LINKED PAGE, ALL	377	244	64.7%	59.8 to 69.4

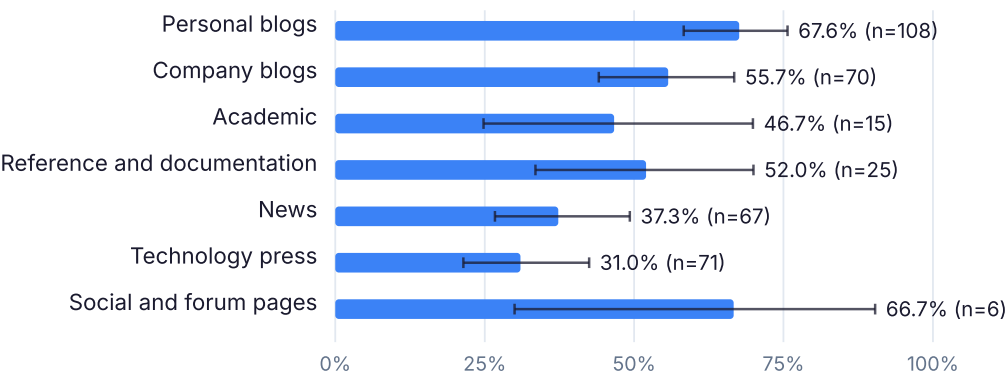
Length drives it. More rules apply to longer text. Unchallenged articles: 29% flagged at 150 to 399 words (only 7 texts), 37% at 400 to 1,199, 64% at 1,200 and over. Comments under 150 words: 0%.

Kind of article matters. Personal blogs and company blogs are flagged most often (next figure).

Period does not. Ordinary comments were flagged 3.4% of the time before ChatGPT and 2.8% in 2025 and 2026.

A few rules cause nearly all the flags. See the rule table below the next figure.

Figure 7. The checker by kind of article. Share of unchallenged articles with at least one category raised.



Genre was assigned to each article's website by a model from what it knows of the site, not by reading the article. 41 unchallenged and 67 challenged articles fall in an unknown group and are left out of this chart.

▼ Data behind this chart

KIND OF ARTICLE	ARTICLES	RAISED A CATEGORY	SHARE	95% INTERVAL
PERSONAL BLOGS	108	73	67.6%	58.3 to 75.7
COMPANY BLOGS	70	39	55.7%	44.1 to 66.8
ACADEMIC	15	7	46.7%	24.8 to 69.9
REFERENCE AND DOCUMENTATION	25	13	52.0%	33.5 to 70.0
NEWS	67	25	37.3%	26.7 to 49.3
TECHNOLOGY PRESS	71	22	31.0%	21.4 to 42.5
SOCIAL AND FORUM PAGES	6	4	66.7%	30.0 to 90.3

RULE	UNCHALLENGED ARTICLES	CHALLENGED ARTICLES	UNCHALLENGED COMMENTS	BOOKS BEFORE 1910
HEAVY EM DASH USE	9.7% of 403	16.4% of 377	0.9% of 108	33 of 300

RULE	UNCHALLENGED ARTICLES	CHALLENGED ARTICLES	UNCHALLENGED COMMENTS	BOOKS BEFORE 1910
RUNS OF NEAR-EQUAL SENTENCE LENGTH	31.0% of 403	43.2% of 377	2.1% of 94	5 of 300
REPEATED SENTENCE OPENINGS	11.9% of 403	22.5% of 377	0.0% of 94	9 of 300
REPEATED PHRASES	18.4% of 403	20.7% of 377	0.0% of 108	0 of 300
LOW SENTENCE-LENGTH VARIATION	0.5% of 403	0.3% of 377	8.5% of 94	11 of 300
FEW CONCRETE DETAILS	0.0% of 403	0.0% of 370	8.9% of 45	9 of 300

Each cell: the share of texts where the rule applied and reached noticeable or high. A rule applies only to texts long enough for it. All other rules fired in under 2% of texts. Challenged articles are the 377 with a linked page. Repeated phrases may partly reflect code and table text that the page scraper kept as prose. The full rule-by-rule table is in [traceability-matrix.csv](#).

What we found wrong with our own checker

The overall flag barely separates challenged from unchallenged articles once the kind of article is held fixed (1.25, interval 0.93 to 1.74). A flag on a long article tells you the article is long and in a common style.

The sentence-variation rule is not specific to the AI era. It fired on 20.3% of the comments written before ChatGPT to which it applies (37 of 182), and on 8.7% and 11.9% in the two later periods.

The dash suggestions often make writing worse (next section), and in old typography they break: where a dash follows another mark, as Victorian books often print it, every offered replacement produced two marks in a row at 45 of 151 dashes (30%). It happened at none of 55 dashes in modern text.

Two lead-in removals can break a sentence. Removing “In addition” from “In addition to the cost, we...” leaves “To the cost, we...”, and removing “It is important to note” from a sentence where the phrase is part of the clause leaves a fragment. Our edit ratings did not hit these, and we found them by testing.

Command-line flags are counted as dashes. A double hyphen in a command such as `git log --online` counts toward the dash rule, and the checker then offers an edit that breaks the command.

Some of the checker page’s own statements go further than the evidence. Its FAQ said readers most often name stock vocabulary, lead-ins, contrast frames, dashes, bullets and bold; accusers mostly named nothing. It said its thresholds were tuned on ten public-domain books so that ordinary prose is not flagged by default, yet those books are flagged 21% of the time. Its dash note cited a rise in published text that no

source supports. Not every flag has a Useful and Not useful button, and we have no usable answers from those buttons yet. The checker page has since been corrected on all of these points.

After this study we changed the checker and its page, in the same release as version 1.1 of this page. The page statements listed above now say what the evidence supports. Dash findings now get an explanation and no replacement, because blind raters preferred the original almost every time. Command-line flags no longer count as dashes. A lead-in is removed only where the removal leaves a sentence, so “In addition to the cost” and “It is important to remember your password” are left alone. The dash note no longer cites a rise in published text, and the dash and “not just X but Y” findings no longer carry a population label. Each fix has a test that fails on the old code. Not yet done: recalibrating the rhythm and repetition rules on a modern sample, and the “I would have changed this anyway” button. Differences for writers who use English as an additional language are unknown, because we had no consented sample. Rules likely to hit those writers harder, and worth testing first, are formal transitions, academic set phrases and sentence-length uniformity. That is a hypothesis from the literature, not a result.

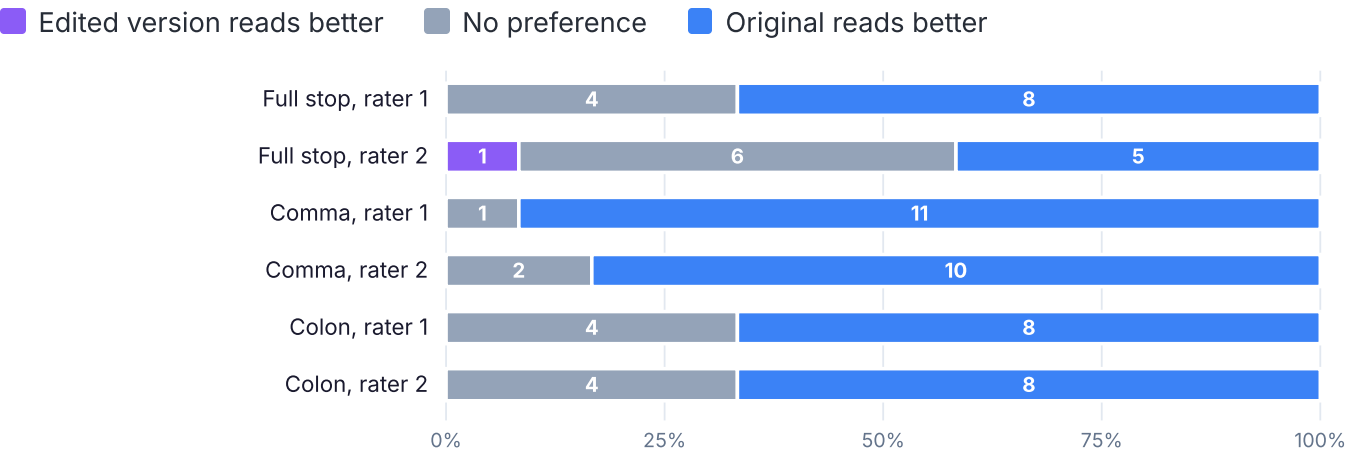
EDITING

Does removing the pattern help?

For dashes, mostly not: replacing flagged dashes made passages read worse more often than better. For stock lead-ins and words the sample was too small to rank.

We applied the checker’s own suggestions to real text and asked two raters, who did not know which version was the original, to compare them for clarity and naturalness, meaning, and errors the edit introduced. The raters were AI models (Claude Sonnet), so this is a model-assisted judgment and not a reading experiment. They agreed moderately (kappa 0.56 on which version reads better in the whole-passage set of modern text, 0.31 on the old-book set, 0.19 to 0.63 across individual options). Hiding which version is A and which is B does not hide which one has the dash.

Figure 8. Does following the checker's dash suggestion improve the passage? Blind comparisons of 12 modern passages per option, by two raters.



Passages come from 12 texts with three options each, so the three options are not independent. Raters were AI models (Claude Sonnet) that did not know which version was the original, which does not hide which version has the dash.

▼ Data behind this chart

OPTION AND RATER	EDITED VERSION BETTER	NO PREFERENCE	ORIGINAL BETTER	EDITED VERSION HAD AN ERROR
FULL STOP, RATER 1	0	4	8	5
FULL STOP, RATER 2	1	6	5	4
COMMA, RATER 1	0	1	11	9
COMMA, RATER 2	0	2	10	8
COLON, RATER 1	0	4	8	8
COLON, RATER 2	0	4	8	4

Every flagged dash replaced, as the checker’s apply button does: 81 passages from 27 texts. The edited version was preferred once by each rater. In the 36 modern passages (from only 12 texts) the original was preferred 27 and 23 times, and the edit was judged to have introduced an error in 22 and 16. A full stop did least harm and a comma the most, with the caveat above.

One dash at a time (209 dash edits): the edited version was preferred once by each rater, out of 209. On modern text the original was preferred in 109 and 143 of 165. This design is harsh, because changing one dash of a pair leaves the other alone.

Removing a stock lead-in or swapping a stock word (12 edits): the edit was preferred 7 and 4 times, the original 2 and 1, and the rest were ties. Too few to rank, and no sign of harm in this sample. The two broken cases above show it is not safe in general.

Raters missed things. They judged meaning unchanged in all but one of 221 single edits, which is not credible where a dash marked an elided name or a sung syllable, and some comma splices passed as grammatical. Read the error counts as lower bounds.

Educated guesses, with the basis stated: a dash that joins two thoughts is often better left alone; when a change is wanted, edit that sentence by hand instead of replacing every dash; the reason accusers cite most, that a text says very little, is not fixed by any punctuation change. Cutting every dash made the passages we tested worse. We did not test cutting contrast frames or lists.

NOT TESTED

What we did not test: cause

Nothing here shows that any pattern makes readers suspicious. That needs an experiment in which only the pattern changes, and we have designed one and not run it.

DESIGN CHOICE	PLAN
STIMULI	24 human-written passages of 90 to 130 words across six genres, each in up to five single-change versions: with the cue, punctuation only changed, structure rewritten, a negative-control change (comma to semicolon), and two cues removed together
READERS	180 adults, each seeing 12 passages in one random version each; non-native English readers analysed as their own group
OUTCOMES, REPORTED SEPARATELY	Perceived likelihood of AI involvement, confidence, naturalness, quality, clarity, and whether the reader would publicly accuse
ANALYSIS	Ordinal mixed models with reader and passage effects; five primary contrasts with Holm correction; a priming condition; effects within 0.2 scale points reported as no practical effect

DESIGN CHOICE	PLAN
NEEDS	Ethical review, a consented panel and a licensed stimulus set including non-native writers. None exists yet, so the experiment has no results

An AI model rating text as “AI-sounding” would not be this experiment: it is not the reader the question is about. The full protocol is in `reading-experiment-protocol.md`. If readers rate dashes as no more AI-like once the rest of the text is fixed, the dash result above is an association with the writers and threads that attract accusation, not with the mark.

METHOD

Method

Finding accusations. 112 phrase searches (for example “reads like AI”, “did you use ChatGPT”) over 46 months of Hacker News through its public search interface, about 55,800 unique comments, then a rule-based filter for replies whose first sentences tie “this” or “you” to an AI-authorship statement. 1,554 comment pairs and 779 article-level pairs were read in full. Quoted lines were removed from every comment before analysis. A case qualifies when the reply asserts, hedges or asks that a specific text is AI-written. One case dates from 2022, so year breakdowns cover 786. Cases: 287 comments; 426 article-level (377 linked pages fetched on October 10, 2026, and 49 posts submitted to Hacker News directly, of which 28 are link submissions whose linked page was not analysed); and 74 on Lemmy, Discourse and Stack Exchange, which are described and not compared because they have no matched controls.

Labels. Model labelers read each pair against fixed written instructions and recorded the target, the strength of the accusation, and the cues the accuser named in their own words. A second labeler agreed on whether a case qualifies in 99% of 240 pairs (kappa 0.97; both are the same model family, so this shows consistency, not a human standard). A model re-reading of 58 accepted cases found 57 correct, and an independent review of 60 random cases found at least one coding error in 10 (see the note under Figure 2). All pattern detection is plain code with exact rules, so a result can be reproduced from the text alone.

Matching. Comments: up to three same-thread comments, same position, similar length, different author, no AI vocabulary in the comment or its replies. Articles: a front-page submission within three days on a different site, similar points and length, with no accusation anywhere in its comments. 229 of 287 comment cases and 260 of 426 article cases found a control.

Statistics. Odds ratios are Mantel-Haenszel across matched sets, with 95% percentile bootstrap intervals over sets. p-values come from a set-level permutation test, and a Benjamini-Hochberg correction covers the patterns tested. A pattern counts as supported only if it was planned in advance, the interval excludes 1, the adjusted intervals for length, AI topic, page structure and exposure agree, and at least 15 matched sets are informative. Patterns found after seeing the data stay at weakly supported. Intervals for shares are

Wilson intervals. The genre comparison pools articles (bootstrap over articles) or keeps the matched sets with same-genre controls (bootstrap over sets).

Re-computation. The matched odds ratios were recomputed from the published per-text pattern flags with code that does not reuse the analysis code, and the figures match to the printed precision; bootstrap intervals differ by up to about 1 unit at the upper end between runs. The flags themselves, the labels and the genre labels are shipped as outputs and were not re-derived. The script is `reproduce.py`.

Baselines and the checker. The comment baselines come from random one-hour windows in three periods, up to 25 comments per window, 40 to 400 words, AI-vocabulary comments removed. The checker is the live engine, run unchanged on every text. The books are Project Gutenberg editions (one, the 1908 Marriott translation of Machiavelli, was published after 1900 and written long before).

Review. Before publication, five independent reviewers (AI models, working from the files and not from our conclusions) checked the method, recomputed the numbers, re-fetched 40 accusations and 20 control comments from Hacker News, read the linguistics and tested the product claims. They found that our first genre comparison dropped the matching, that the “em dash” flag mixed two marks, that 28 article cases analysed the wrong text, and that several sentences claimed more than the data show. Each was corrected on this page. A sixth review covered the page itself.

LIMITATIONS

Limitations

The publisher is in the sample. UltraTextGen made the checker whose false alarms this study measures, and we report where it falls short.

One site, English only. Hacker News is a technical, English-language audience. Results do not transfer to other communities or languages. Reddit is not covered.

Unchallenged is not human. Some unchallenged texts may be AI-written and some challenged ones may be human. The comparison is between texts readers challenged and texts they did not.

We cannot compute how often a pattern draws an accusation. We found accusations by searching for accusation phrases. We have no count of texts that were read and let pass.

Exposure. Hacker News does not publish how many people read a comment. A comment nobody read cannot be accused. Matching reduces this and does not remove it.

Labels are model-made and imperfect. Both labelers, the genre labels and the edit raters are AI models. About one case in six had a coding error in an independent re-check. To see whether this matters, two further blind AI codings of all 787 cases were made. They rejected 19 and 17 cases as not qualifying (23 by either, agreement kappa 0.72). The share naming no reason came out at 55.5% and 65.8% against 57.3% originally, so treat the 57% as a range of about 55% to 66%. The em dash was named 18 or 19 times by every coding. Dropping the 23 rejected cases, the matched odds ratio for the em dash character in short comments is 9.1 (4.55 to 24.6) against 8.4, and for runs of short sentences 6.4. Dropping the doubtful ones

as well (606 cases) it is 9.7 and 5.3. The conclusions do not depend on the coding errors we found; the exact counts do. These codings are the same kind of model work, not human review. See `recoding-comparison.csv`.

What we cannot verify. Deleted, flagged and removed replies are invisible, so whether a comment was challenged, and who answered first, depends on what remained visible. The data files hold no post identifiers, so an outside reader cannot check that any single case exists, only recompute the statistics. The search service caps results at 1,000 per query, which limited recall from February 2026.

Article text is today's version for 377 cases, which may differ from what the accuser read.

Small periods. The 2023 and 2024 comment results rest on 43 matched sets and 4 to 8 informative ones.

Genre was assigned by a model from the website name, about half with low confidence, with an unknown group of 108 articles.

The edit ratings come from AI models, from 12 modern texts, agree only moderately, and are lower bounds on errors.

Most baselines are not known to be human either. Only the books and the 2019 to 2021 comments are certain to predate AI.

No non-native writers, no reader experiment, no causal claim.

Sentence rules depend on a heuristic sentence splitter, which in scraped articles can follow page layout instead of prose.

What we do about it: every number here can be recomputed from the data files, the checker's own "Useful / Not useful" answers are the first-party test of any rule we change, and the reading experiment is the next study. The checker corrections above shipped with this version, and we still have no first-party answers from the checker's feedback buttons to test against.

DATA

Data, files and reproducibility

The files contain counts and pattern flags only. They hold no usernames, no links to individual posts and no text from the accused or comparison writing, so that no commenter is pointed out.

FILE	CONTENTS	SIZE
ai-writing-footprint-study.pdf	The full study, printed from this page	586 KB
accusation-cases-deidentified.csv	Start here: one row per analysed accusation (787), with year, kind of text, what the accuser named, other readers' replies and the author's reply. No text, no link, no name	142 KB

FILE	CONTENTS	SIZE
text-features.csv	One row per analysed text (1,754): whether each pattern is present and what each checker rule rated it, and whether the text is a linked page or a post submitted to the site directly. The input to reproduce.py	262 KB
pattern-associations.csv	Every pattern, in comments and articles: counts, matched odds ratio, intervals, adjusted ratios, tests and the evidence status	11 KB
named-cues.csv	What accusers named, by pattern and kind of text, with 95% intervals	3 KB
pattern-definitions.csv	The 30 patterns: definition, exact matching rule and where it misfires	7 KB
traceability-matrix.csv	Each of the checker's 22 rules: what the data show, how often it fired on unchallenged text and on books before 1910, and the action we would take	10 KB
checker-trigger-rates.csv	How often each rule fired, by kind of text, length band and year	101 KB
genre-adjustment.csv	Odds ratios before and after holding the kind of article fixed	1 KB
ordinary-comments-features.csv	One row per ordinary comment (4,934) drawn at random from 2019 to 2021, 2023 to mid 2024 and 2025 to 2026: dash character, pattern flags and checker rule levels	449 KB
hn-period-baseline-summary.csv	Pattern shares and checker flags in those comments, by period and length, with intervals; the no-ai-vocab copy leaves out comments that mention AI	19 KB
hn-period-baseline-summary-no-ai-vocab.csv	The same without comments that mention AI	19 KB
dash-variants.json	Odds ratios for the em dash character, the typed double hyphen and either, by period	3 KB
edit-ratings.csv	The blind ratings of the checker's edit suggestions, one row per rater and edit, with the option and which version was preferred. No text	58 KB

FILE	CONTENTS	SIZE
recoding-comparison.csv	The second and third blind codings of the 787 accusations, and the matched odds ratios with contested cases dropped. Counts only	5 KB
reading-experiment-protocol.md	The reading experiment we designed and did not run	7 KB
reproduce.py	Recomputes the matched odds ratios from text-features.csv	3 KB

To reproduce the odds ratios, put `reproduce.py` and `text-features.csv` in one folder and run `python3 reproduce.py text-features.csv` (Python 3.9 or later, standard library only). It prints the matched odds ratios for the em dash and for runs of short sentences.

To check your own writing, open the AI Writing Footprint Checker.

CITE

How to cite this study

UltraTextGen Research. "What Makes Writing Sound AI Generated? A Study of Public Accusations, Writing Patterns and False Alarms." Version 1.1, October 13, 2026.
<https://ultratextgen.com/research/ai-writing-footprint-study/>

Please cite the version number. A later version will keep this address and say what changed.

Version 1.1 (October 13, 2026): added a second, independent coding of all 787 accusations and a check that the findings survive it (see Limitations); corrected the statements about the checker, most of which have now been fixed in the checker and its page (see "What we found wrong with our own checker"); added `recoding-comparison.csv`.

SOURCES

External sources

That detectors misjudge writing by non-native English speakers is shown by Liang, Yuksekgonul, Mao, Wu and Zou in GPT detectors are biased against non-native English writers

(<https://arxiv.org/abs/2304.02819>) (Patterns, 2023), which is why we make no claim for these writers and why the checker does not score authorship.

That some words rose sharply in biomedical abstracts after 2022, across a whole body of papers and not for any single document, is reported by Kobak, Gonzalez-Marquez, Horvat and Lause in Delving into LLM-assisted writing in biomedical publications through excess vocabulary

(<https://arxiv.org/abs/2406.07016>) (Science Advances, 2025).

That untrained readers judge machine-written text at or near chance, and rely on flawed shortcuts, is shown by Jakesch, Hancock and Naaman in Human heuristics for AI-generated language are flawed (<https://arxiv.org/abs/2206.07271>) (PNAS, 2023; six experiments, 4,600 participants), by Clark and colleagues in All That's 'Human' Is Not Gold (<https://arxiv.org/abs/2107.00061>) (arXiv preprint, 2021), and by Fleckenstein and colleagues in Do teachers spot AI? (<https://doi.org/10.1016/j.caeai.2024.100209>) (Computers and Education: Artificial Intelligence, 2024), where neither novice nor experienced teachers could pick out ChatGPT texts.

One study reports an exception: Russell, Karpinska and Iyyer, People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text (<https://arxiv.org/abs/2501.15654>) (ACL 2025), found that the majority vote of five annotators who use ChatGPT heavily misclassified only 1 of 300 articles. Some accusations in our data were surely right.

The books used as a pre-2022 baseline are public-domain editions from Project Gutenberg (<https://www.gutenberg.org/>). The Hacker News comments were read through the site's public search interface (<https://hn.algolia.com/api>).