Do AI Models Speak Human?

TextPulse Research. Working Paper.

Abstract

Users who want an AI model to write like a person usually assume that the machine register is a default that an instruction can lift. We tested that assumption. Ten passages of published academic prose from PubMed Central, all from articles deposited before 2022, gave ten topics. Four flagship models (GPT-5.6 Sol, Claude Opus 5, Gemini 3.7 Flash and DeepSeek V4 Pro) wrote about 300 words on each topic under three prompts: a plain request; a detailed brief on the properties that separate human academic prose from assistant prose (uneven sentence length, plain words, no triads, no connective openers, hedging, asides); and the same brief with a human passage as an example. Every text was placed on the frozen 47-feature stylometric human-to-AI spectrum from our earlier work and scored by GPTZero, alongside the ten human passages. The brief worked on the surface: the pooled spectrum position fell from 148 under the plain prompt to 30 under the styled one (human passages 10), 14 of the 40 styled texts landed inside the middle half of the human anchors, and Claude Opus 5 moved past the human median. The models overshot every property the brief named, writing more unevenly and at a lower reading grade than the journals, and left the one property it did not name, vocabulary range, where tuning had put it. GPTZero did not move: all 120 AI texts, under every prompt and from every model, received a document AI probability of 1.00, and all ten human passages received 0.00. The human exemplar added nothing. One pass through TextPulse moved the same texts a mean of 47 points toward the human end while keeping 97 percent of the meaning, and moved the detector where instruction had not. We read the result through the published measurements of preference tuning: supervised fine-tuning and RLHF narrow the distribution a model samples from, an instruction is a condition on that narrowed distribution rather than a replacement for it, and a probability-based detector reads the process, not the surface. Prompting cannot undo preference tuning. All texts, prompts, detector replies, scores and code are released.

1. Introduction

Ask a current AI model to write a paragraph of academic prose and the result is fluent, well organised and, to a detector or to a trained reader, recognisably machine written. The usual reply from users is that the model was not told to write like a person. Prompt guides, forum threads and paid “humanizing prompts” all rest on the same assumption: that the machine register is a default the model falls into, and that an explicit instruction, or an example of human writing, will lift it out. This study tests that assumption directly.

We took ten passages of published academic prose from PubMed Central, all from articles published before 2022, so none can carry text from the ChatGPT era. For each passage we wrote a topic statement, then asked four flagship models, GPT-5.6 Sol, Claude Opus 5, Gemini 3.7 Flash and DeepSeek V4 Pro, to write about 300 words on that topic under three prompts. The first prompt asked only for a research passage. The second added detailed instructions on the properties that separate human academic writing from assistant writing: uneven sentence length, ordinary word choice, no triads, no connective openers, hedging, asides, small irregularities. The third added a human passage on a different topic as an example to imitate. Every output was placed on the stylometric human-to-AI spectrum built in our earlier work and scored by GPTZero, and the ten human passages were scored the same way. Finally, every AI output was passed once through TextPulse, so the reader can see what a purpose-built rewrite does to the same text.

The question has a practical edge. If instruction alone could make a model write like a person, AI detection would already be finished as a technology, and humanizers would be redundant. If it cannot, then the machine register is not a default the model chooses but a property the model carries, and the reason lies in how these models are trained. Section 2 reviews that training, from supervised instruction tuning to reinforcement learning from human feedback and its successors, and what the published work says it does to output diversity. Sections 4 to 6 report the measurements. Section 7 gives the verdict.

2. Background and related work

2.1 What separates human from machine prose

The stylometric differences between human and model text are well documented. Muñoz-Ortiz et al. (2023) found that human news text has a more scattered sentence-length distribution, shorter constituents and a wider emotional range than LLaMA output, which used more numbers, auxiliaries and pronouns. Herbold et al. (2023) compared 1,000 student essays with ChatGPT essays and found the model text more uniform in structure and rated higher by teachers. Reviriego et al. (2024) measured lower lexical diversity in ChatGPT-3.5 than in humans on the same tasks, with GPT-4 closing the gap on vocabulary but not on the underlying distribution. Park et al. (2025) tested the latent community structure of paraphrased text and rejected the hypothesis that model text and human text share a distribution. Our own series built a 47-feature stylometric fingerprint (TextPulse Research, 2026e), a classifier and spectrum from it (2026f), a 1,057-word lexicon of vocabulary that models inject and suppress (2026d), and a burstiness study showing that sentence-length variation is the single largest separating feature (2026b). Kobak et al. (2025) tracked the model vocabulary into the literature itself: across 15 million PubMed abstracts, words such as delve, underscore, crucial and pivotal rose abruptly after late 2022, and at least 13.5 percent of 2024 abstracts showed signs of model processing. Liang et al. (2024) found the same words in 6.5 to 16.9 percent of AI-conference peer reviews.

2.2 Can a model be told to write like a person

Fewer studies test whether instruction removes those differences. Perkins et al. (2024) used prompt engineering to make GPT-4 submissions evade Turnitin and found that adding burstiness and misspellings lowered detector accuracy by about 17 percentage points, a gain that came from breaking the text rather than from writing better. Shi et al. (2024) passed detectors with instructional prompts and word substitution. Matsubara (2025) had ChatGPT rewrite an abstract in the author's own style and found the result classified as machine text by every detector, with a median AI probability of 100 percent, while the author's genuine abstract scored a median of 0 percent. Wang et al. (2025) tested whether current models can imitate the implicit style of ordinary authors from examples and concluded that they cannot yet: the outputs matched content and surface register but not the authors' stylometric profile. Xu et al. (2026) found that raw base-model output, before preference tuning, already reads as human to GPTZero and Pangram, which points at the tuning rather than the pretraining as the source of the machine register. None of these studies gives the model the full stylometric brief and a human exemplar, measures the result on a fixed spectrum, and compares it with human prose on the same topics. That is the gap this paper fills.

2.3 How the models are trained, and what it does to their writing

Every model in this study is a pretrained language model that has been through several stages of post-training. Pretraining fits the model to the distribution of a large text corpus; a base model at this stage samples from something close to that distribution, with all its variety. The first post-training stage is supervised fine-tuning on demonstrations written or selected by human labellers, which teaches the model the assistant format. The second is preference tuning. In the original recipe, reinforcement learning from human feedback (RLHF), labellers rank pairs of model outputs, a reward model is trained to predict those rankings, and the language model is optimised against the reward model with a penalty for drifting from the supervised model (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022). Constitutional AI replaced part of the human labelling with model-written critiques against a written set of principles (Bai et al., 2022). Direct preference optimisation removed the separate reward model but kept the objective (Rafailov et al., 2023). Every current flagship uses some version of this pipeline, and several add reinforcement learning on verifiable tasks.

The published effect of preference tuning on the text is consistent across studies. Kirk et al. (2024) measured each stage of the RLHF pipeline and found that RLHF improved generalisation but sharply reduced output diversity, on every diversity measure they used, compared with supervised fine-tuning alone. O'Mahony et al. (2024) attributed this mode collapse to specific fine-tuning components and to bias in the preference data. Padmakumar and He (2024) had people write essays with a base model, with a preference-tuned model, and alone, and found that only the preference-tuned model reduced the diversity of what the group wrote, because its suggestions were less varied than the base model's. Singhal et al. (2023) found that most of the reward gained

in RLHF was explained by response length alone. Yun et al. (2025) showed that the chat template itself, the role markers and special tokens of an instruction-tuned model, collapses output diversity even at high sampling temperature. Mahapatra (2026) probed seventeen models and found that instruction-tuned systems collapse entropy along discourse and structural dimensions regardless of scale, and Gude et al. (2026) compared two generations of models and found the more aligned generation less diverse in grammar and lexicon. Janus (2022) first documented the phenomenon in text-davinci-002, which OpenAI later confirmed was tuned on high-rated demonstrations rather than by RLHF proper, which shows that the narrowing begins with preference data of any kind, not with the reinforcement step specifically.

The mechanism is not mysterious. A reward model trained on human rankings learns what raters prefer on average: complete sentences of even length, clear signposting, balanced lists, confident summary, a polished lexicon. Optimising against that average pushes every output toward it. Human academic prose is not the average of what raters prefer; it is the output of individual writers under deadline, with uneven sentences, plain words, stray asides and unfinished thoughts. A model tuned toward the rater average can be told to imitate that texture, and it will make gestures toward it, but the instruction is applied on top of a distribution that has already been narrowed. The question for this study is how far those gestures reach.

3. Data and methods

3.1 Human passages

The human baseline is ten passages from the PubMed Central open-access subset, drawn from the human side of the TextPulse master corpus (TextPulse Research, 2026e). Only articles with a PMC identifier below 8,700,000 were eligible, which restricts the pool to articles deposited before 2022, before any public chat model existed. Each passage is a contiguous excerpt of one article chunk, cut at a sentence boundary at about 300 words, and copied without any edit. Chunks with extraction artefacts (glued table captions, missing spaces) and narrative or dialogue passages were skipped by rule, and the ten were drawn with a fixed random seed from the remainder: four in linguistics, three in law, three in the humanities. The selection script is released with the data. Table 1 lists the passages.

Passage	PMC article	Discipline	Words	Spectrum	GPTZero AI
H01	PMC7810638	Linguistics	293	4	0.00
H02	PMC8292947	Linguistics	300	-37	0.00
H03	PMC8147217	Linguistics	312	19	0.00
H04	PMC8484525	Linguistics	326	-20	0.00
H05	PMC8600356	Law	292	5	0.00

Passage	PMC article	Discipline	Words	Spectrum	GPTZero AI
H06	PMC8345368	Law	306	46	0.00
H07	PMC8056094	Law	328	20	0.00
H08	PMC8576790	Humanities	311	7	0.00
H09	PMC7585998	Humanities	309	15	0.00
H10	PMC8655481	Humanities	299	40	0.00

Table 1. The ten human passages. Spectrum is the position on the frozen human-to-AI axis (human anchor median 0, AI anchor median 100); GPTZero AI is the document AI probability.

3.2 Topics and prompts

For each human passage we wrote a one-sentence topic statement describing its subject in neutral terms. The four models then wrote on that topic under three prompts. The plain prompt (C1) asked for a passage of about 300 words for the body of a research article, in continuous prose, with no title, headings or lists. The styled prompt (C2) added an explicit brief drawn from the stylometric literature and from our own fingerprint study: vary sentence length a great deal, mixing sentences under ten words with sentences over thirty-five; prefer plain words over polished ones and avoid a named list of model vocabulary (delve, leverage, crucial, pivotal, multifaceted, nuanced, landscape, underscore, tapestry, robust, foster, comprehensive, intricate, notably, highlight); do not group ideas in threes; do not open sentences with Moreover, Furthermore, Additionally, Overall or In conclusion; do not end with a summary sentence; hedge and judge the way a researcher does; name specific studies, places, years or numbers; use parenthetical asides; allow an occasional fragment, a repeated word and an uneven paragraph; use no em dashes. The exemplar prompt (C3) added the styled brief plus one of the human passages, on a different topic, with the instruction to match its rhythm, word choice and level of formality without reusing its content. The exact prompt text is released in `prompts.json`.

3.3 Models

The four models are the current flagship on each provider's API at the time of the run (29 August 2026): gpt-5.6-sol (OpenAI), claude-opus-5 (Anthropic), gemini-3.7-flash (Google) and deepseek-v4-pro (DeepSeek). Each cell of the design is a separate, stateless request at the provider's default temperature. Extended reasoning was switched off where the provider allows it, as in our humanizer comparison, since the question is what the model writes rather than how long it deliberates. A reply outside 200 to 420 words was regenerated once; both attempts are kept in the release. Word counts of the outputs averaged 341 against 308 for the human passages.

3.4 TextPulse pass

Every AI output was submitted once to TextPulse's production engine under the settings a signed-in user gets by default: Academic mode, intensity 0.5, American English, em dashes stripped. One submission per text, no reruns, output saved verbatim. This arm is reported separately in Section 6 and is not part of the answer to the main question.

3.5 Measures

Spectrum position. Every text is placed on the frozen human-to-AI spectrum built for the humanizer comparison (TextPulse Research, 2026g), which is a logistic classifier over 47 stylometric features trained on 121,000 paired human and AI texts from our earlier fingerprint study, projected onto the axis from the human centroid to the AI centroid and scaled so that the human anchor median reads 0 and the AI anchor median reads 100. The middle half of the human anchors lies between -36 and 33; the middle half of the AI anchors between 61 and 139. Nothing was refit for this paper, so the numbers are directly comparable with the humanizer paper. The classifier's probability that a text is AI written (p_ai) is reported alongside the axis position.

Detector. Each text was sent once to GPTZero's document endpoint (v2 API). We report the document AI probability and the share of texts classified as human. The raw replies are cached and released.

Style components. To show what moves the spectrum, we report six of its inputs directly: the coefficient of variation of sentence length (burstiness), MATTR-50 lexical diversity, mean word length, Flesch-Kincaid grade, the density of AI-leaning words per 1,000 from the frozen 1,057-word lexicon, and the share of sentences opening with a connective.

TextPulse arm. For each humanized text we also report the BGE document cosine to its AI source and the share of words changed, so the shift on the spectrum can be read against what was kept.

3.6 Statistics

Group means are reported with 95 percent bootstrap confidence intervals (5,000 resamples, seed 12). With ten texts per model and prompt, the intervals are wide and are drawn on every figure; the paper reports differences of group location, not significance tests between individual cells. The TextPulse shift is computed per text as humanized minus source and summarised the same way. Everything is scripted; `score.py` produces `per-text-scores.csv` and `analysis-results.json`, and `figures.py` draws every figure from those files.

4. The spectrum moves when the model is told to move it

Figure F1 and Table 2 give the main result on the stylometric spectrum. The ten human passages sit at a mean of 10 (95 percent CI -5 to 24), inside the middle half of the human anchors, which is where published academic prose should sit. Under the plain prompt the four models wrote at a mean of 148 (95 percent CI 127 to 168), past the AI anchor median, with GPT-5.6 Sol at 185 and

Gemini 3.7 Flash at 222, further from human prose than the median AI text in the 121,000-text anchor set. Claude Opus 5 (79) and DeepSeek V4 Pro (105) were closer but still well outside the human band. Only one of the forty plain-prompt texts fell inside the human middle half.

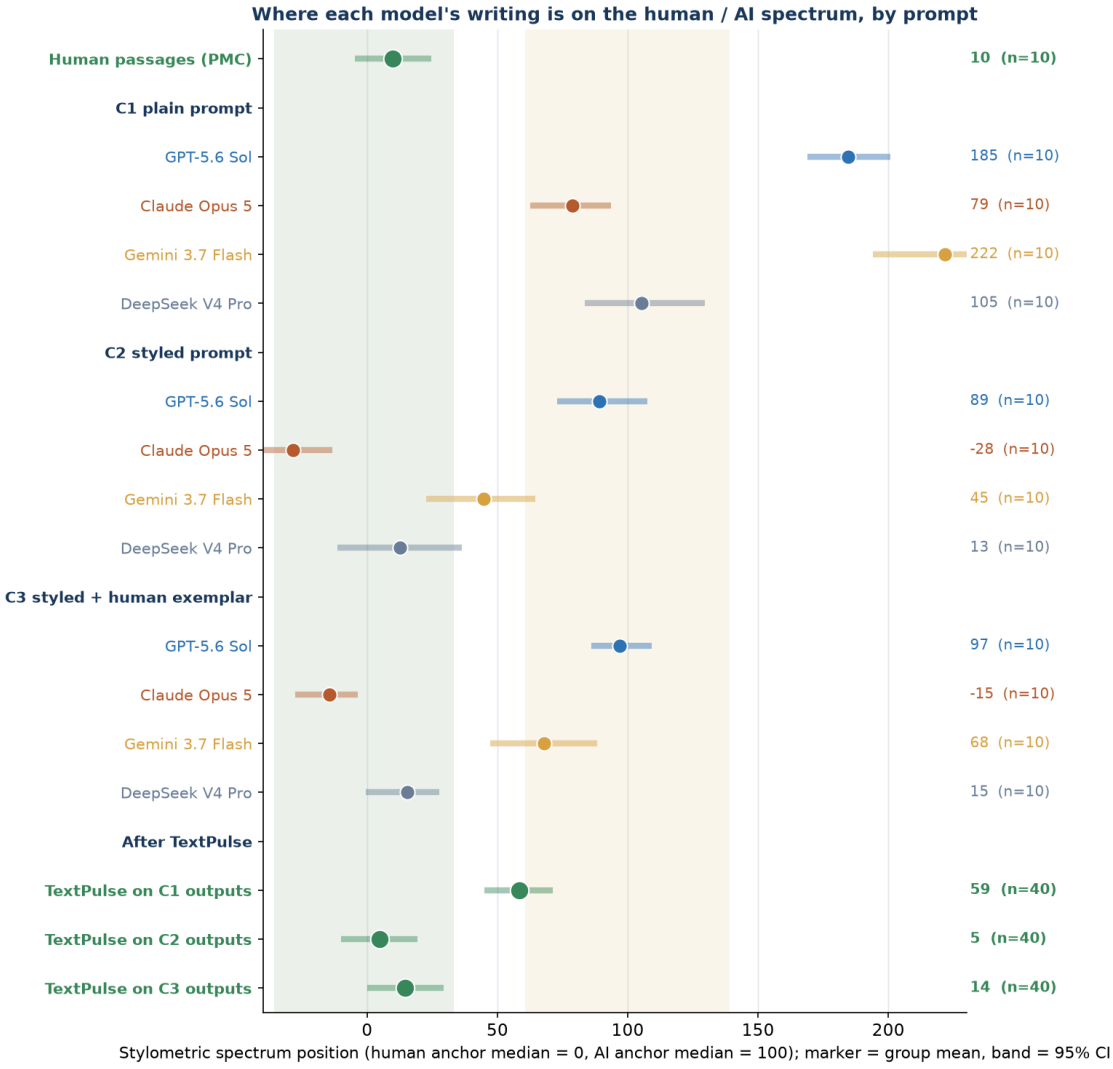

Figure F1. Spectrum position by model and prompt, with the ten human passages and the TextPulse pass. Marker = group mean, band = 95 percent bootstrap CI; green shading = middle half of the human anchors, amber = middle half of the AI anchors.

The styled prompt changed this a great deal. The pooled mean fell from 148 to 30 (95 percent CI 14 to 47), and 14 of the forty styled texts landed inside the human middle half. Claude Opus 5 moved furthest, from 79 to -28, past the human median and to the human side of every one of the ten PMC passages' mean; DeepSeek V4 Pro moved to 13, Gemini 3.7 Flash to 45, and GPT-5.6 Sol, the least responsive, to 89, still inside the AI band. Adding a human exemplar on top of the styled brief (C3) did not help further: the pooled mean was 41 (95 percent CI 27 to 57), slightly further from human prose than C2 for every model except Claude's neighbour DeepSeek, and 18 texts fell inside the human band. The ordering of the models was the same under every prompt: Claude, then DeepSeek, then Gemini, then GPT.

Table 3 and Figure F4 show which components moved. The styled brief asked for uneven sentence length, and the models delivered it beyond the human level: sentence-length CV rose from 0.32 under the plain prompt to 0.58 under the styled one, against 0.4 in the human passages. Mean word length fell from 6.45 to 5.39 characters, below the human 5.67, and the Flesch-Kincaid grade fell from 21.4 to 13.8, far below the human 20.4: told to write plainly, the models wrote more plainly than the journals do. AI-lexicon density fell from 137 to 78 words per 1,000, below the human 107, and connective sentence openers, which the brief banned by name, disappeared entirely (0.00) where the human passages use them in 0.03 of sentences. One component did not move at all: lexical diversity (MATTR-50) was 0.88 under the plain prompt and 0.877 under the styled one, against 0.804 for the human passages. The brief said nothing about vocabulary range, and the models' vocabulary range stayed where preference tuning left it, wider than a human researcher's.

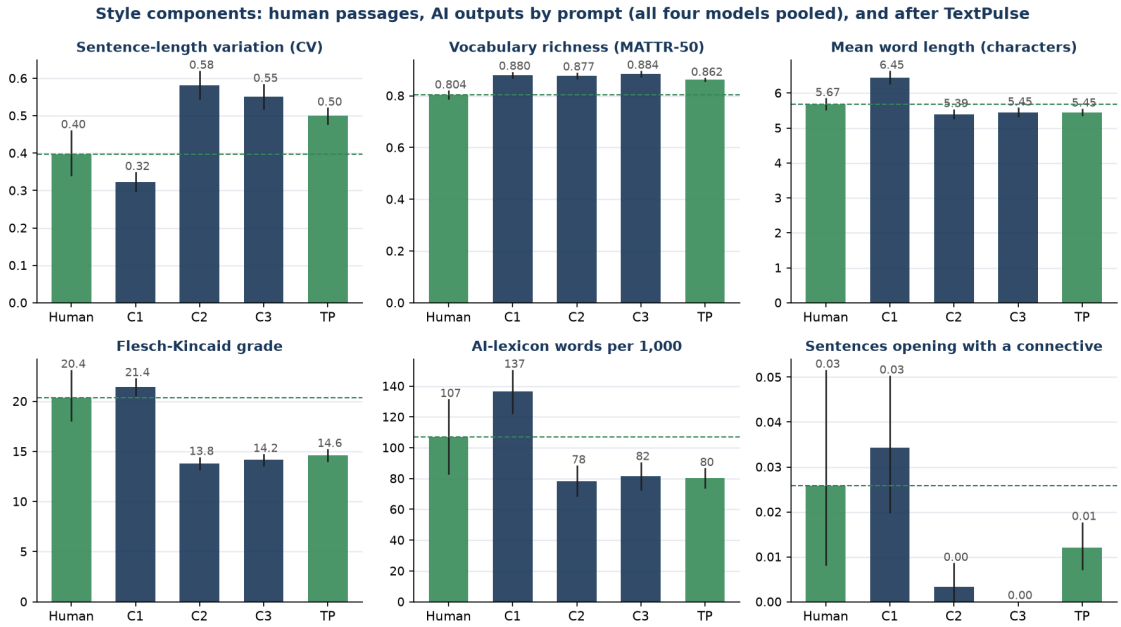

Figure F4. Style components: human passages, AI outputs by prompt with the four models pooled, and the TextPulse pass. Dashed line = human mean.

The pattern is that of a model executing a list. Every property named in the brief moved, several of them past the human value, and the property the brief did not name stayed put. The result is text that scores as human on a 47-feature spectrum because the features the brief happened to cover dominate that spectrum, without being human in the way the spectrum was designed to detect.

5. GPTZero is not moved at all

Figure F2 gives the detector result, and it needs one sentence. All 120 AI texts, under every prompt and from every model, received a GPTZero document AI probability of 1.00 and were classified as AI; the lowest probability among the 120 was 1.00. All ten human passages received a probability of

0.00 and were classified as human; the highest was 0.0001. The styled prompt, which moved the spectrum from 148 to 30 and put 14 texts inside the human band, moved GPTZero by nothing. Claude Opus 5's styled texts, which sit on the human side of the human median on the spectrum, were flagged with the same certainty as Gemini's plain-prompt texts at 222.

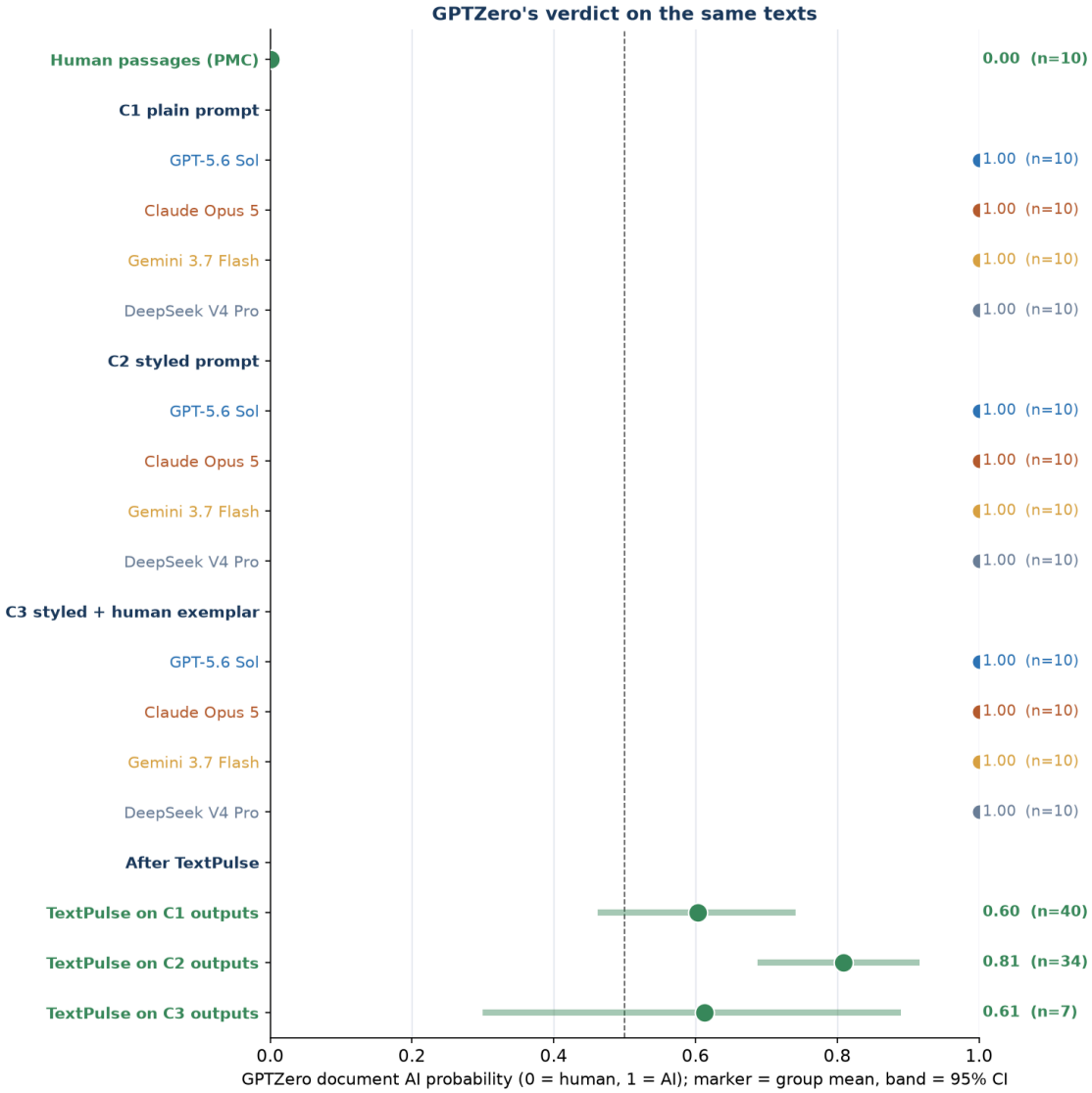

Figure F2. GPTZero document AI probability by model and prompt. Every AI cell sits at 1.00; the human passages at 0.00.

Figure F3 puts the two measures together. The horizontal axis is the stylometric spectrum, the vertical axis the detector. The human passages sit at the bottom left. Every AI text, wherever it sits on the horizontal axis, sits on the top line. Instruction moved the texts along the horizontal axis and left the vertical axis untouched.

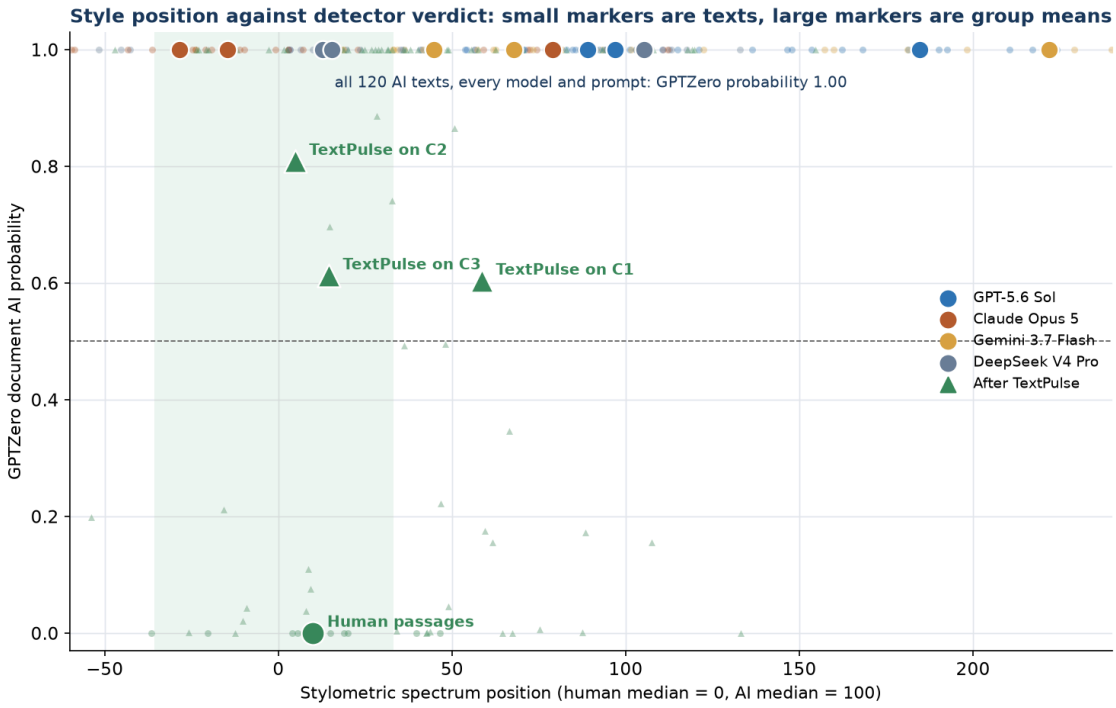

Figure F3. Spectrum position against GPTZero probability. Small markers are texts, large markers are group means. All 120 AI texts sit at probability 1.00 regardless of spectrum position.

The two instruments disagree because they read different things. The spectrum reads 47 surface counts: sentence lengths, word lengths, punctuation, openers, vocabulary lists. A model can be told what those counts should be and can hit them. GPTZero reads the probability of each token under a language model and the way that probability varies across the text. A preference-tuned model asked to write “unevenly” produces an uneven text whose every token is still the token such a model would produce; the unevenness is itself generated in the model’s own manner. The detector sees the generation process, not the surface it was told to imitate. This is the distinction between a text that has the statistics of human writing and a text that came from the human distribution, and it is why Sadasivan et al. (2025) tie detectability to the distance between the two distributions rather than to any list of features.

6. What one pass through TextPulse does

Every AI text was passed once through TextPulse. Figure F5 and Table 4 show the shift on the spectrum. Across the 120 texts the mean shift was -47 points toward the human end (95 percent CI -55 to -39), with the largest shifts on the texts that started furthest away: Gemini’s plain-prompt texts moved from 222 to 80 and GPT’s from 185 to 95. The pooled plain-prompt texts moved from 148 to 59, the styled texts from 30 to 5, and the exemplar texts from 41 to 14. After the pass, 54 of the 120 texts sat inside the human middle half (12 of the plain-prompt texts, 22 of the styled, 20 of the exemplar texts), against 33 before. The rewrite kept the text: document cosine to the source

averaged 0.972 while 49 percent of the words changed. AI-lexicon density fell by a further 19 words per 1,000 and the reading grade by 1.8; sentence-length variation was left where the source had it.

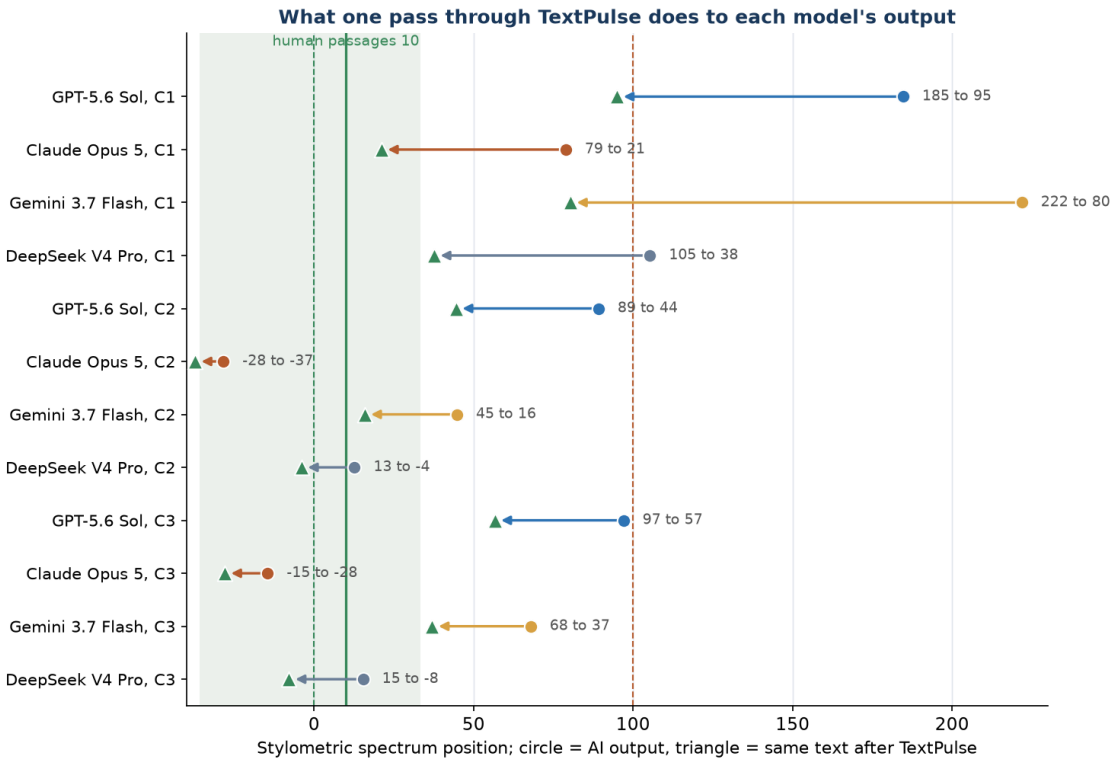

Figure F5. The TextPulse shift on the spectrum, per model and prompt. Circle = AI output, triangle = the same text after one pass; solid green line = the human passages.

GPTZero was run on 81 of the 120 humanized texts (all 40 plain-prompt texts, 34 styled and 7 exemplar texts) before the detector budget for the study was closed; the remaining 39 have spectrum scores only. Of the 81, 25 were classified as human: 17 of 40 plain-prompt texts, 6 of 34 styled texts and 2 of 7 exemplar texts. The mean probability fell from 1.00 to 0.60 on the plain-prompt texts and 0.81 on the styled ones. Two points follow. First, a rewriting engine that changes the token sequence moves the detector where an instruction to the generating model does not, because the rewrite changes the process that produced the text rather than the surface the model was asked to produce. Second, the styled texts were harder to move than the plain ones: the source already carried the surface markers of human writing, so the rewrite had less to change and the detector kept more of its signal. Instruction and humanization are not additive. The reader should note that these 81 detector results describe TextPulse's own engine and are reported for completeness; the humanizer comparison paper (TextPulse Research, 2026g) is the controlled test of that engine against its competitors.

7. Verdict: prompting cannot undo preference tuning

The question in the title has a two-part answer. On the surface, yes: told which properties separate human prose from machine prose, current flagship models can reproduce those properties, and on a spectrum built from those properties they can be placed among human texts. Underneath, no: a detector that reads the generation process rather than the surface classifies every one of the 120 texts as machine written with full confidence, whatever the prompt, whatever the model, and however human the surface reads. An example of human writing in the prompt does nothing that the written brief did not already do.

The reason is in Section 2.3. Every model in this study has been through supervised fine-tuning and preference optimisation, and the published measurements of those stages agree that they narrow the distribution the model samples from (Kirk et al., 2024; O'Mahony et al., 2024; Padmakumar and He, 2024; Yun et al., 2025; Mahapatra, 2026). The narrowing is not a style the model adopts and can drop on request. It is the shape of the distribution after training, and an instruction is a condition applied to that distribution, not a replacement for it. Asked to write unevenly, the model samples an uneven text from the narrowed distribution. The surface counts change; the per-token signature that a probability-based detector reads does not, because the same narrowed process produced every token. This is also why the components moved in the way they did in Section 4: the brief named sentence length, word choice and openers, and each was pushed past the human value, since a preference-tuned model executes an instruction to the letter; the brief did not name vocabulary range, and vocabulary range did not move. A human writer does not need to be told these things and does not overshoot them, because a human writer is not applying a list to a narrowed distribution. Xu et al. (2026) report that raw base-model output, before any preference tuning, reads as human to the same detectors. Base models are not offered to the public as writing assistants. The models that are offered carry the tuning, and the tuning is what the detector reads.

Two practical conclusions follow. For a writer, no prompt tested here, including the most detailed brief we could write plus a human exemplar, produced a text that a university-grade detector would treat as human, and the prompt that moved the stylometric surface furthest produced text that read simpler and more uneven than the journals it was imitating. For the detection question, the result cuts both ways. Surface stylometry, including our own spectrum, can be steered by instruction and should not be used on its own as evidence of authorship; the probability-based detector was not steered at all in this study, but it was steered by a rewriting engine in Section 6, which is the finding of the humanizer literature reviewed in our earlier paper. The stable fact underneath both results is that a preference-tuned model, asked to write like a person, writes like a preference-tuned model asked to write like a person.

8. Limitations

Ten topics and four models give ten texts per cell, so the confidence intervals in Table 2 are wide and cell-level differences of a few points should not be read. The human baseline is ten passages from three disciplines, all drawn from the same corpus family that supplied the spectrum's human

anchors, so the human passages are expected to sit near the human median by construction; the comparison that matters is between the AI cells and that fixed reference. The spectrum is a linear model over 47 surface features and Section 4 shows that it can be steered, which is a finding of this paper as well as a limitation of the instrument. GPTZero is one detector; we did not test Turnitin, Originality or Pangram on these texts, and a detector built on different principles could respond to the styled prompt differently. Reasoning was switched off in every model, and a reasoning pass might change how a model applies a style brief. The attribution of the detector's stability to preference tuning rests on the literature in Section 2.3 rather than on a base-model arm in this study; the design deliberately includes no untuned model. The TextPulse arm is a description of our own engine and has the conflict of interest that implies; it is reported with its full data and is not the basis of the paper's conclusion. Finally, the study is a snapshot of four models on one day; the models change and so does the detector.

Prompt	Model	n	Spectrum mean	95% CI	p_ai	GPTZero AI	Classified human
Human passages		10	9.9	-4.8 to 24.5	0.18	0.00	10 of 10
C1	GPT-5.6 Sol	10	184.6	168.9 to 200.7	0.98	1.00	0 of 10
C1	Claude Opus 5	10	78.9	62.5 to 93.7	0.66	1.00	0 of 10
C1	Gemini 3.7 Flash	10	221.8	194.1 to 248.1	0.99	1.00	0 of 10
C1	DeepSeek V4 Pro	10	105.2	83.6 to 129.6	0.75	1.00	0 of 10
C1	All four	40	147.6	127.0 to 168.4	0.85	1.00	0 of 40
C2	GPT-5.6 Sol	10	89.1	72.9 to 107.6	0.69	1.00	0 of 10
C2	Claude Opus 5	10	-28.5	-44.9 to -13.2	0.03	1.00	0 of 10
C2	Gemini 3.7 Flash	10	44.8	22.7 to 64.6	0.19	1.00	0 of 10
C2	DeepSeek V4 Pro	10	12.6	-11.5 to 36.4	0.22	1.00	0 of 10

Prompt	Model	n	Spectrum mean	95% CI	p_ai	GPTZero AI	Classified human
C2	All four	40	29.5	13.6 to 46.6	0.28	1.00	0 of 40
C3	GPT-5.6 Sol	10	97.0	85.9 to 109.2	0.76	1.00	0 of 10
C3	Claude Opus 5	10	-14.6	-27.7 to -3.5	0.05	1.00	0 of 10
C3	Gemini 3.7 Flash	10	67.9	47.2 to 88.2	0.35	1.00	0 of 10
C3	DeepSeek V4 Pro	10	15.3	-0.5 to 27.7	0.21	1.00	0 of 10
C3	All four	40	41.4	26.7 to 56.9	0.34	1.00	0 of 40

Table 2. Spectrum position, classifier probability (p_ai) and GPTZero document probability by prompt and model. C1 plain prompt, C2 styled prompt, C3 styled prompt plus a human exemplar.

Component	Human	C1 plain	C2 styled	C3 exemplar	After TextPulse
Sentence-length CV	0.40	0.32	0.58	0.55	0.50
MATTR-50	0.804	0.880	0.877	0.884	0.862
Mean word length	5.67	6.45	5.39	5.45	5.45
Flesch-Kincaid grade	20.4	21.4	13.8	14.2	14.6
AI-lexicon words per 1,000	107	137	78	82	80
Connective openers (share of sentences)	0.03	0.03	0.00	0.00	0.01

Table 3. Style components, group means. The four models are pooled within each prompt (n = 40 per prompt, 10 human, 120 after TextPulse).

Prompt	Model	Spectrum before	Spectrum after	Cosine to source	Words changed	GPTZero AI after	Classified human after
C1	GPT-5.6 Sol	185	95	0.968	59%	1.00	0 of 10
C1	Claude Opus 5	79	21	0.962	60%	0.50	5 of 10
C1	Gemini 3.7 Flash	222	80	0.957	70%	0.37	7 of 10
C1	DeepSeek V4 Pro	105	38	0.961	58%	0.54	5 of 10
C2	GPT-5.6 Sol	89	44	0.979	49%	0.90	1 of 10
C2	Claude Opus 5	-28	-37	0.974	37%	0.73	1 of 4
C2	Gemini 3.7 Flash	45	16	0.978	44%	0.77	2 of 10
C2	DeepSeek V4 Pro	13	-4	0.971	44%	0.79	2 of 10
C3	GPT-5.6 Sol	97	57	0.981	48%	0.45	2 of 3
C3	Claude Opus 5	-15	-28	0.974	39%	not scored	not scored
C3	Gemini 3.7 Flash	68	37	0.984	42%	0.54	0 of 2
C3	DeepSeek V4 Pro	15	-8	0.975	42%	0.92	0 of 2

Table 4. One pass through TextPulse. Spectrum before and after are the cell means; cosine is the BGE document similarity between the humanized text and its AI source; GPTZero columns cover the 81 humanized texts that were scored (see Section 6).

Data availability

The ten human passages, every AI output including rejected attempts, every TextPulse output, the exact prompts, the cached GPTZero replies, the per-text scores, the analysis results and every script are released with this paper. The archive is deposited on Zenodo (DOI 10.5281/zenodo.22159495).

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., … & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073. https://arxiv.org/abs/2212.08073

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1706.03741

Gude, A., Santos-Rios, R., Bond, F., Flickinger, D., Gomez-Rodriguez, C., & Zamaraeva, O. (2026). More aligned, less diverse? Analyzing the grammar and lexicon of two generations of LLMs. arXiv:2605.06030. https://arxiv.org/abs/2605.06030

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

Janus. (2022). Mysteries of mode collapse. LessWrong. https://www.lesswrong.com/posts/t9svvNPNmFf5Qa3TA

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., & Raileanu, R. (2024). Understanding the effects of RLHF on LLM generalisation and diversity. International Conference on Learning Representations. https://arxiv.org/abs/2310.06452

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

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., … & Zou, J. Y. (2024). Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. Proceedings of the 41st International Conference on Machine Learning. https://arxiv.org/abs/2403.07183

Mahapatra, R. (2026). From context shift to stylistic collapse: Why training objectives matter more than scale. arXiv:2605.28826. https://arxiv.org/abs/2605.28826

Matsubara, S. (2025). Large language model content detectors: Effects of human editing and ChatGPT mimicking individual style, a preliminary study. Journal of Investigative Medicine, 73(8), 666-672. https://doi.org/10.1177/10815589251352572

Munoz-Ortiz, A., Gomez-Rodriguez, C., & Vilares, D. (2023). Contrasting linguistic patterns in human and LLM-generated text. arXiv:2308.09067. https://arxiv.org/abs/2308.09067

O'Mahony, L., Grinsztajn, L., Schoelkopf, H., & Biderman, S. (2024). Attributing mode collapse in the fine-tuning of large language models. ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models. https://openreview.net/pdf?id=3pDMYjpOxk

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., … & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35. https://arxiv.org/abs/2203.02155

Padmakumar, V., & He, H. (2024). Does writing with language models reduce content diversity? International Conference on Learning Representations. https://arxiv.org/abs/2309.05196

Park, M., Choi, Y., & Jeon, J.-J. (2025). Does a large language model really speak in human-like language? arXiv:2501.01273. https://arxiv.org/abs/2501.01273

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

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36. https://arxiv.org/abs/2305.18290

Reviriego, P., Conde, J., Merino-Gomez, E., Martinez, G., & Hernandez, J. A. (2024). Playing with words: Comparing the vocabulary and lexical diversity of ChatGPT and humans. Machine Learning with Applications, 18, 100602. https://doi.org/10.1016/j.mlwa.2024.100602

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

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

Singhal, P., Goyal, T., Xu, J., & Durrett, G. (2023). A long way to go: Investigating length correlations in RLHF. arXiv:2310.03716. https://arxiv.org/abs/2310.03716

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., … & Christiano, P. (2020). Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33. https://arxiv.org/abs/2009.01325

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

TextPulse Research. (2026b). Sentence-length burstiness as a cross-disciplinary and cross-model signal of AI rewriting. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22033063

TextPulse Research. (2026d). The vocabulary fingerprint of AI rewriting: A 1,057-word lexicon from 60,786 paired academic texts. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22028377

TextPulse Research. (2026e). Stylometric fingerprints of AI rewriting: Punctuation, syntax, and model attribution. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22040684

TextPulse Research. (2026f). Human versus AI text classification from stylometric features across 121,092 academic texts. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22048477

TextPulse Research. (2026g). A controlled comparison of AI text humanizers on academic writing. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22157928

Wang, Z., Tripto, N. I., Park, S., Li, Z., & Zhou, J. (2025). Catch me if you can? Not yet: LLMs still struggle to imitate the implicit writing styles of everyday authors. Findings of the Association for Computational Linguistics: EMNLP 2025. https://arxiv.org/abs/2509.14543

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

Yun, L., An, C., Wang, Z., Peng, L., & Shang, J. (2025). The price of format: Diversity collapse in LLMs. arXiv:2505.18949. https://arxiv.org/abs/2505.18949