
Burstiness Was Measured Wrong, and Prompting Cannot Aim It

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Burstiness is a signal people already use to judge machine text, and its operational
2 definition in practice is an unreviewed detector heuristic confounded with length:
3 raw sentence-length variance grows with the mean. We treat sentence boundaries
4 as a point process in word-time, scored with the Goh–Barabási coefficient under
5 the Kim–Jo finite- n correction: bounded and scale-free, no model in the loop,
6 screened by a Goodhart guard and a coherence gate. The protocol is paired evaluation
7 on 200 era-, topic-, and register-matched Wikinews articles dated 2016–2021:
8 five prompting modes, nine model families, three seeds, uniform normalization of
9 generated text before scoring. Scored raw, markdown scaffolding manufactures
10 dispersion and produces a parity verdict that normalization reverses (instruction-
11 mode 0.005, $p = 0.10$, becomes a significant 0.028 deficit, $p = 3 \times 10^{-5}$), and the
12 uninstructed dispersion gap more than doubles (0.043 to 0.097). Of the forty-five
13 family-by-mode cells, thirty-two land significantly flatter than the human articles,
14 eight show no significant difference, and five land significantly more dispersed:
15 machine flatness is a strong central tendency that prompting moves without target-
16 ing. The same instruction leaves four families short, lands two on the human level
17 (CI-bounded equivalence), moves one further away, pushes two past, and few-shot
18 exemplars of typical human rhythm are the flattest mode by rank-biserial in seven
19 of nine families. We model human rhythm; we do not build detector evasion.

20 1 Introduction

21 Burstiness is the variation in sentence length that gives human prose its rhythm and is one of the
22 signals most often cited to separate human writing from machine writing [Fraser et al., 2024]: while
23 human authors usually alternate long and short clauses, language models tend toward a flatter middle.
24 The deployed operational definition traces to an unreviewed detector technical note [Fraser et al.,
25 2024, Radvand et al., 2025], and its statistic, raw sentence-length variance, is confounded with
26 absolute length: uniformly long sentences can score burstier than genuinely varied short ones.

27 Sentence-length dispersion has been a stylometric variable since Yule [1939], Williams [1940], but
28 as a signal for authorship it has never been given a reproducible definition that divides rhythm apart
29 from length. The aim of this paper is to give this definition and use it to benchmark the prompting
30 techniques people already rely on, modeling human rhythmic patterns rather than building a tool
31 whose purpose is to evade detection.

32 **A definition.** We treat sentence boundaries as a point process in word-time and score the sentence-
33 length sequence with the Goh–Barabási burstiness coefficient [Goh and Barabási, 2008] under the
34 Kim–Jo finite- n correction [Kim and Jo, 2016], written B_n . It is bounded in $[-1, 1]$, needs no
35 reference model, and depends on the coefficient of variation alone, being scale-free: it measures
36 rhythm relative to typical sentence length instead of conflating the two. Alongside B_n we report

a Gini robustness check [Gini, 1912], Miller–Madow-corrected punctuation entropy as a separate segmentation axis, and a memory coefficient M for lag-1 order; surprisal, the strong but model-relative detector signal, stays deliberately outside the construct. Because a metric that becomes a target invites gaming, we pair B_n with two safeguards: a Goodhart guard that rejects degenerate constructions that inflate any dispersion score, and a coherence gate [Shypula et al., 2025] that keeps incoherent high-variance text from winning the metric.

A benchmark of five prompting techniques. We evaluate control (a bare brief), instruction, few-shot human exemplars [Brown et al., 2020, Min et al., 2022], an outline-then-write plan mode [Wang et al., 2023], and a draft-critique-rewrite self-refine loop [Madaan et al., 2023] against 200 English Wikinews articles [Wikinews, 2021] dated 2016–2021, with era, topic, and register held fixed. Each generation is length-matched to its paired article, and every cell is sampled at three seeds. Generated text is uniformly normalized before scoring, because markdown scaffolding and echoed headlines in raw output manufacture dispersion the human baseline lacks, and on the audited leg this choice more than doubles the uninstructed dispersion gap. The identical design runs across nine model families, spanning open-weight checkpoints (Qwen3.6-27B, Gemma 3 27B), API-served models (DeepSeek V4 Pro, GLM-5.2, Kimi K2.6, Grok 4.3), and closed frontier systems (Grok 4.5, Gemini 3.1 Pro, Claude Opus 4.8).

The aim problem. Prompting changes rhythm, and the landing is family-dependent. Uninstructed, eight of the nine families write significantly flatter than the human articles, with control deficits 0.027 to 0.128; Grok 4.5’s control-mode difference is not significant (-0.009 , $p = 0.086$). The three largest gaps belong to frontier flagships, so scale does not close the gap. The identical instruction leaves four significantly short of the human dispersion level, lands two on it, moves one further away, and pushes two significantly past it, and a family’s uninstructed gap does not predict where it lands. Few-shot exemplars of typical human rhythm are the flattest mode by rank-biserial in seven of nine families, and within the Grok lineage one generation widens the gap under instruction while the next inverts past the human level.

Contributions.

1. A reproducible definition of burstiness that is model-free and scale-free, hardened against gaming: the Goh–Barabási coefficient under the Kim–Jo finite- n correction, companions (Gini check, segmentation-diversity axis, memory coefficient; surprisal kept outside) and two anti-gaming safeguards (Section 3).
2. A paired protocol on era-, topic-, and register-matched news articles: uniform normalization before scoring, seed-averaging that avoids pseudo-replication, within-family multiplicity correction, parity only as CI-bounded equivalence with pre-specified margins, a contamination probe, and a moderation-refusal audit (Sections 4.1–4.3).
3. The aim problem across nine families, five modes, three seeds: identical instruction leaves four significantly short of the human dispersion level, lands two on it, moves one further away, and pushes two significantly past it, and few-shot exemplars are the flattest mode by rank-biserial in seven of nine families (Section 4).

2 Related work

2.1 Burstiness in detection

The operational definition of burstiness in current practice originates in a detector technical note (Tian/GPTZero), which pairs perplexity with a single burstiness scalar [Fraser et al., 2024, Ravand et al., 2025]. It is widely used and not peer-reviewed, the formalization gap Section 3 closes. Tarm and Onan [2025] compare diffusion and autoregressive text stylometrically with burstiness among the features. DivEye [Basani and Chen, 2025] tracks how unpredictability fluctuates across a text, the nearest text-side analog to timing variation. Surveys [Fraser et al., 2024] treat perplexity-burstiness as a standard pair and document detector vulnerability to paraphrasing.

85 2.2 Stylistic variation and diversity

86 The stylistic-variation benchmark [Milika et al., 2025] evaluates frontier models across prompt set-
87 tings and scores how varied the output is, with no human reference distribution to score against.
88 Effective semantic diversity [Shypula et al., 2025] gates diversity on quality, and we adopt that
89 method as the composite’s coherence gate so incoherent high-variance text is not rewarded.

90 3 Formalizing burstiness

91 We replace the field’s unreviewed operational definition (Section 2) with a model-free, reproducible
92 measurement instrument.

93 3.1 Burstiness as a point process

94 Let a document D be segmented into sentences with word-length sequence $L = (l_1, \dots, l_n)$. We
95 treat the sentence boundaries as a point process in word-time, so L is a sequence of inter-event
96 intervals, and score it with the Goh–Barabási burstiness coefficient [Goh and Barabási, 2008],

$$B = \frac{\sigma - \mu}{\sigma + \mu},$$

97 where μ and σ are the mean and standard deviation of L . B is bounded in $[-1, 1]$: $B = -1$ for a
98 metronomic sequence, $B = 0$ for a Poisson one, and $B \rightarrow 1$ for a maximally bursty one. Because
99 real documents give short interval sequences, we apply the Kim–Jo finite- n correction [Kim and Jo,
100 2016], writing B_n for the corrected coefficient.

101 B_n is *scale-free*: as a function of the coefficient of variation $r = \sigma/\mu$ alone, $B = (r - 1)/(r + 1)$,
102 it measures variation *relative* to typical sentence length, so rhythm is not conflated with length.
103 Classically, B_n is a bounded reparametrization of the coefficient of variation, weak alone in the
104 stylometric literature [Yule, 1939, Williams, 1940]. What is new is the model-free formulation
105 hardened against gaming, its paired evaluation on a matched modern corpus, and the nine-family
106 prompt benchmark built on it (Section 4).

107 3.2 The construct and its companions

108 Burstiness proper is the single scale-free coefficient B_n . We report four companions alongside it,
109 none folded into the construct:

- 110 • **Gini robustness check.** The Gini coefficient [Gini, 1912] of the sentence lengths is a scale-
111 free dispersion measure that aggregates over all pairwise differences, so it is less hostage
112 to a single long sentence. It agrees with B_n in direction on every cell of the nine-family
113 benchmark where either difference is significant (Section 4); agreement shows a result is
114 not an estimator artifact.
- 115 • **Segmentation diversity, a separate axis.** The Miller–Madow-corrected Shannon entropy
116 of the punctuation distribution measures how varied the segmentation symbols are, a dis-
117 tinct axis from rhythm. Typographic quotes are normalized to ASCII uniformly across all
118 nine legs, so smart-quote typography cannot register as diversity.
- 119 • **Memory, the order term.** The Goh–Barabási memory coefficient M [Goh and Barabási,
120 2008], the lag-1 correlation between consecutive interval lengths, is the order information a
121 dispersion statistic discards. In Section 4 three families’ few-shot legs overshoot the human
122 articles on memory while undershooting on dispersion.
- 123 • **Predictability, kept outside.** Mean per-token surprisal and the mean absolute consecu-
124 tive difference, under a fixed reference model, form the strong but model-relative detector
125 signal. The definition places them *outside* the construct: they are a first-moment quantity,
126 whereas burstiness is a second moment of a derived sequence. The benchmark scores the
127 four model-free dimensions only and reports no surprisal values.

128 We compute raw sentence-length variance $\text{var}(L)$ only as a diagnostic; it is confounded with abso-
129 lute length by construction, the central defect of the colloquial statistic.

3.3 Anti-gaming screens and reproducibility

A metric that becomes a selection score invites gaming. Degenerate texts can reach a high-dispersion score (one giant sentence among fragments, mechanical alternation, punctuation spam). A Goodhart guard flags these; the full flag list ships in the reference implementation (Appendix C). We count B_n only above a reliability floor of ten sentences. Re-punctuating a fixed text can move B_n across much of its bounded range, so the structural metric must never select generations alone; following the quality-and-diversity argument [Shypula et al., 2025], the composite pairs it with a coherence gate, so incoherent high-variance text cannot win the metric. The benchmark of Section 4 is descriptive and selects nothing by the metric; neither the guard nor the gate filters the scored generations here.

The metric, the guard, and the multiscale profile the guard’s wash-out check reads are pure standard library in `experiments/burstiness_pp.py`, with a self-check; Appendix C reproduces the full implementation. They run without a model; of the companions above, only surprisal would need one, one more reason it stays outside the construct.

4 The aim problem: how reliably instructions move the construct

How much of the human-machine rhythm gap prompting closes, and how predictably, is the question this paper answers. The same wording scatters the nine families, and this section measures that scatter.

4.1 Corpus and generation

The corpus is 200 English Wikinews articles [Wikinews, 2021] (CC BY 2.5), dated 2016–2021 and therefore pre-model, with era, topic, and register held fixed, so within a pair only authorship differs. The human side of each pair is the plain article body, regenerable from the source Wikinews wikitext by a deterministic parse (`experiments/ablation/clean_wikinews.py`) that, on a held-out check of 100 articles, reproduces the shipped human bodies exactly. The first leg is a frozen open-weights Qwen3.6-27B (bf16, temperature 0.7, top- p 0.95, reasoning mode enabled, reasoning stripped before scoring). Eight further families join in Section 4.4. For each headline the model writes a news article from a brief containing only the headline and the paired human article’s word count, so each generation targets its pair’s length. We compare five non-overlapping prompting modes: a *control* brief with no stylistic instruction, an *instruction* mode that adds a metric-free persona directive to write as a professional reporter in a natural human voice, never flat or mechanical, restated near the end of the prompt, *few-shot* exemplars of matched human news prose [Brown et al., 2020, Min et al., 2022], an outline-then-write *plan* mode [Wang et al., 2023], and a draft-critique-rewrite *self-refine* loop [Madaan et al., 2023]. The three few-shot exemplars are representative human news reports, matched by register and topic and centered on the median of the human B_n distribution (each near $B_n = -0.416$). The selection uses B_n only to find the exemplars nearest that median, never the high-dispersion end, so the few-shot mode sees typical human rhythm. No prompt names the measured axis, so no mode teaches to the test. We sample each (mode, headline) cell at three fixed seeds. Of the leg’s 3,000 cells, 2,996 completed; we excluded the four that exceeded the reasoning budget rather than scoring their incomplete output. We average affected headlines over their remaining seeds, so every mode retains 200 matched pairs. The entire run is scripted and reproducible from the released artifacts.¹ Appendix A lists the verbatim wording of every mode and the rule that composes them.

4.2 Normalization: stripping the wrapper before scoring

Model output does not arrive as plain prose. A generation-hygiene guard flagged markdown scaffolding and echoed headline blocks in every family we benchmark, from 23–38% of control-mode generations on one leg (DeepSeek V4 Pro: 23% markdown, 38% title lines) to 100% on another (Gemma 3 27B markdown; headline echoes on the Grok 4.5 and Gemini 3.1 Pro legs). All of it is formatting the plain-prose human baseline lacks. A segmenter fed raw output scores a markdown header or an

¹ Scoring code, prompts, the metric implementation, per-family scored evaluations (normalized protocol, plus the audited leg’s raw-protocol archive), and generation metadata accompany the paper as supplementary material. The full anonymized repository, adding raw generations for eight of nine legs and the research database, is at <https://anonymous.4open.science/r/burstiness-was-measured-wrong-1p4fm>. De-anonymized links appear in the camera-ready.

177 echoed dateline as a short pseudo-sentence, and those fragments manufacture length dispersion and
 178 segmentation symbols the article prose does not have. Every generated leg is therefore normalized
 179 before scoring: a deterministic script (`experiments/ablation/normalize_generation.py`)
 180 drops the echoed headline block and unwraps markdown syntax while preserving the article prose;
 181 the human baseline is untouched. A word-loss audit of the full first leg finds median 0.0%, mean
 182 1.7%, and at most 7.4% of a generation lost to normalization, that maximum being the headline
 183 block itself.

184 Scaffolding sits outside the construct: B_n is defined over the sentence lengths of prose, and a mark-
 185 down header or an echoed dateline is not a sentence of the article. Scoring it credits formatting
 186 as rhythm, the same class of inflation the Goodhart guard rejects (Section 3), so normalization is
 187 a corpus-level Goodhart correction, applied uniformly to every leg. Its size is measurable on the
 188 audited leg (Qwen3.6-27B): scored raw, instruction-mode B_n shows parity (0.005, $p = 0.10$) where
 189 normalized text shows a significant deficit (0.028, $p = 3 \times 10^{-5}$); raw control-mode gaps on segmen-
 190 tation entropy and memory (0.038 and -0.060 , both significant) vanish on normalized text (-0.001
 191 and -0.007), which is why we describe that leg’s control gap as confined to the dispersion dimen-
 192 sions; and the control-mode dispersion gap on normalized text is more than twice the raw estimate
 193 (0.043 to 0.097), with per-seed rank-biserial rising from 0.27–0.34 to 0.61–0.64. Raw scorings are
 194 archived with the released artifacts.

195 4.3 Analysis

196 Because the three seeds of a cell share their human counterpart, pooling them as separate pairs
 197 would pseudo-replicate the human documents and overstate significance. We therefore average each
 198 headline’s metric vector across its seeds first and run one Wilcoxon signed-rank test per mode and
 199 dimension over the 200 independent matched pairs. Matched rank-biserial is reported as effect size.
 200 Each test compares one mode against the human articles; a family’s narrowing or widening from
 201 one mode to another compares two estimates and is not itself a tested contrast.

202 Every family ran the identical twenty-cell grid (five modes by four dimensions), corrected within
 203 family with Benjamini–Hochberg at $q = 0.05$, Holm reported alongside; the correction script and
 204 each leg’s census ship with the released artifacts. Across the nine grids, 124 of the 180 cells are
 205 significant at the per-cell 0.05 level; 120 survive Benjamini–Hochberg and 103 survive Holm. All
 206 four cells that fall under Benjamini–Hochberg sit on companion dimensions, none on B_n ; the one
 207 the prose discusses, Qwen3.6-27B’s instruction-mode memory overshoot ($p = 0.045$), we mark
 208 exploratory and do not build on. No correction is applied across families; Table 1 carries per-cell
 209 stars with a pointer to each family’s corrected census in Appendix B.

210 Matching claims are never stated as failures to reject. A match requires the 95% CI of the median
 211 paired delta inside a pre-specified margin; margins 0.03, 0.05, and 0.08 were fixed before the cross-
 212 family runs. We also report m^* , the smallest margin the CI would clear, the larger of the two absolute
 213 CI bounds. CIs on the median paired delta are percentile bootstrap intervals: the 2.5th and 97.5th
 214 percentiles of the medians of 2,000 resamples of the 200 matched pairs. Per-seed B_n effect sizes
 215 released alongside are consistent with the seed-averaged results.

216 4.4 One unified instruction for 9 families

217 The design of Section 4.1 runs identically on every family: the same 200 headlines, the same
 218 five wordings, three seeds, uniform normalization, the same paired scoring. The nine families are
 219 Qwen3.6-27B and Gemma 3 27B (open weights, bf16, temperature 0.7, top- p 0.95); DeepSeek V4
 220 Pro, Grok 4.3, GLM-5.2, and Kimi K2.6 (API-served, provider-default decoding; GLM-5.2 ignores
 221 a requested temperature and Kimi K2.6 rejects one); and Grok 4.5, Gemini 3.1 Pro, and Claude Opus
 222 4.8 (closed APIs; Grok 4.5 exposes neither temperature nor top- p , Gemini 3.1 Pro ran at tempera-
 223 ture 1.0 with provider-managed thinking, and Claude Opus 4.8 exposes no temperature). Thinking
 224 modes, effort levels, and output caps are recorded in each leg’s released metadata. Anthropic bills
 225 reasoning against the output cap, so the shared 16,000-token cap that leaves other families room
 226 to write leaves Claude Opus 4.8 none at maximum effort: the self-refine rewrite spends the whole
 227 budget reasoning and emits no prose at all. We ran that leg at high effort with a 32,000-token cap,
 228 which restores a prose budget comparable to the other families and matches the thinking level Gem-
 229 ini 3.1 Pro ran at. The leg completed all 3,000 cells with no truncated or empty output. Decoding is

Table 1: The cross-family aim problem on dispersion: median $\Delta B_n = \text{human} - \text{model}$ per mode, seed-averaged Wilcoxon paired tests over the same 200 matched headlines per family. Positive places the model below the human articles (flatter), negative above them (more dispersed). Matched rank-biserial in parentheses; * $p < 0.05$, ** $p < 0.01$ per cell; each family’s corrected twenty-cell grid is in Appendix B, with no correction across families (Section 4.3). The two rows below the rule are successive Grok generations: instruction widens the gap at 4.3 and inverts past the human level at 4.5.

Model	Control	Instruction	Few-shot	Plan	Self-refine
Gemini 3.1 Pro	0.128 (0.85)**	0.071 (0.56)**	0.169 (0.94)**	0.070 (0.58)**	0.039 (0.27)**
GLM-5.2	0.126 (0.85)**	0.061 (0.57)**	0.148 (0.94)**	0.080 (0.65)**	0.011 (0.13)
Claude Opus 4.8	0.119 (0.80)**	−0.037 (−0.25)**	0.039 (0.34)**	−0.000 (0.01)	−0.065 (−0.50)**
Qwen3.6-27B	0.097 (0.67)**	0.028 (0.34)**	0.119 (0.80)**	0.045 (0.41)**	0.039 (0.26)**
Kimi K2.6	0.095 (0.69)**	−0.006 (−0.00)	0.123 (0.83)**	0.026 (0.24)**	−0.070 (−0.57)**
Gemma 3 27B	0.069 (0.58)**	0.049 (0.44)**	0.063 (0.65)**	0.061 (0.57)**	0.029 (0.30)**
DeepSeek V4 Pro	0.035 (0.34)**	0.000 (0.05)	0.029 (0.24)**	0.013 (0.15)	−0.018 (−0.10)
Grok 4.3	0.027 (0.27)**	0.050 (0.42)**	0.108 (0.76)**	0.061 (0.56)**	0.019 (0.19)*
Grok 4.5	−0.009 (−0.14)	−0.025 (−0.20)*	0.077 (0.69)**	−0.009 (−0.02)	−0.059 (−0.41)**

not matched across providers, so the benchmark reads as nine separate paired comparisons against the same human baseline, each family compared against the human articles alone. Where a provider returns reasoning content separately we strip it before scoring. Gemini’s thoughts are never returned, and Gemma 3 has no thinking mode. We benchmark single generators only; multi-agent systems are excluded because they break the one-generator pairing.

Completeness is disclosed per leg. Qwen3.6-27B excludes four cells for exceeded reasoning budgets (Section 4.1), and Grok 4.5 reports two empty self-refine outputs, truncations by signature, seed-averaged away. Provider content moderation removes cells on two legs, by two different mechanisms. Moonshot refuses explicitly: 25 of Kimi K2.6’s 3,000 cells return HTTP 400, one Hong Kong protest headline refused at every mode and seed. Zhipu refuses silently: 36 of GLM-5.2’s cells return an ordinary HTTP 200 with empty output, concentrated on the same cluster of politically sensitive headlines. The harness records both as empty cells and seed-averaging excludes them, so every paired statistic is computed over the headlines the model actually wrote; a benchmark that does not audit empty outputs would silently absorb the second mechanism into its scores. The remaining five legs generated all 200 headlines in every mode.

Table 1 reports the B_n cell for every mode on every family, Figure 1 plots all forty-five cells, and Appendix B carries each family’s full four-dimension grid with its correction census.

Uninstructed, eight of nine families write significantly flatter than the human articles, with control deficits from 0.027 to 0.128. Grok 4.5’s control-mode difference is not significant (−0.009, $p = 0.086$). The three largest gaps belong to the frontier flagships, Gemini 3.1 Pro (0.128, rank-biserial 0.85), GLM-5.2 (0.126, 0.85), and Claude Opus 4.8 (0.119, 0.80), which start within 0.009 of one another, so frontier scale alone does not close the gap. Deficits of at most 0.128 on a coefficient bounded in $[-1, 1]$ are consistent signatures, well short of the gross flatness the detector literature describes [Fraser et al., 2024].

On four families the instruction narrows the gap without closing it: Gemini 3.1 Pro lands at 0.071, GLM-5.2 at 0.061, Qwen3.6-27B at 0.028, and Gemma 3 27B at 0.049.

Two families land on the human dispersion level, and we state both as CI-bounded equivalence with pre-specified margins. DeepSeek V4 Pro closes a significant control gap (0.035, rank-biserial 0.34, $p = 3 \times 10^{-5}$) to a median delta of 0.000 with 95% CI $[-0.012, 0.026]$ ($p = 0.57$), inside the tightest margin, $m^* = 0.026$. Kimi K2.6 closes 0.095 (0.69) to −0.006 with CI $[-0.034, 0.020]$ ($p = 0.96$), bounded at the middle margin, $m^* = 0.034$. The DeepSeek landing is one-dimensional: its instructed leg overshoots human segmentation entropy (−0.145, rank-biserial −0.47, $p < 10^{-5}$) while matching dispersion. Kimi’s instructed leg shows no per-cell significant difference on any dimension, though only its dispersion cell carries an equivalence bound.

Five of the forty-five cells land significantly past the human articles. Claude Opus 4.8 and Grok 4.5 cross under instruction (−0.037, rank-biserial −0.25, $p = 0.002$, surviving both corrections;

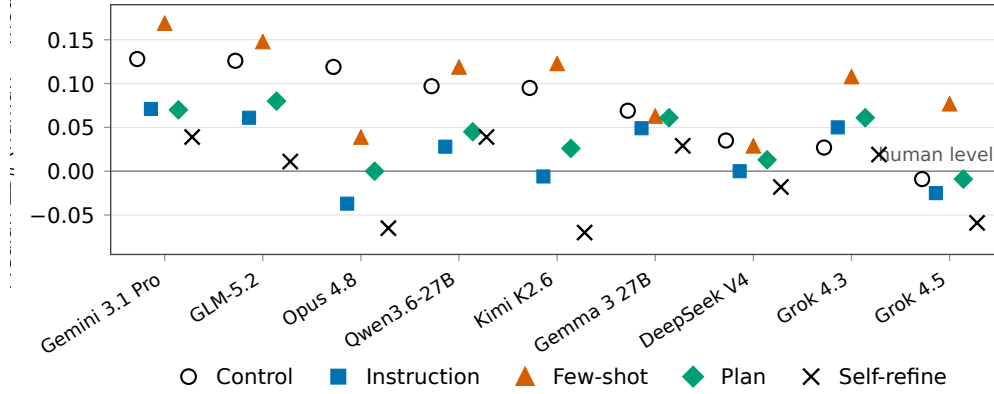


Figure 1: Median $\Delta B_n = \text{human} - \text{model}$ for every family and prompting mode, the forty-five cells of Table 1. Reading key: the gray line is the human dispersion level; markers above it are flatter than the human articles, markers below it more dispersed. Families are ordered by their control-mode gap. Every control marker except Grok 4.5’s sits above the line, few-shot sits highest in six of nine families by median, and is the flattest mode by rank-biserial in seven: the difference is Gemma 3 27B, where few-shot carries the larger effect size but control the larger median, and instruction lands on both sides of the line, including across the three leftmost families whose control markers nearly coincide. Per-cell significance is in Table 1.

266 -0.025 , -0.20 , $p = 0.015$, surviving Benjamini–Hochberg, falling under Holm), and self-refine
 267 pushes three families past (Kimi K2.6 -0.070 , Claude Opus 4.8 -0.065 , Grok 4.5 -0.059 , all
 268 surviving both corrections). Instruction widens Grok 4.3’s gap from 0.027 to 0.050 (the modes’
 269 CIs overlap, $[0.005, 0.044]$ and $[0.030, 0.068]$), and Grok 4.5, one generation later, inverts past the
 270 human level. The flatness tell is not monotone even within a single vendor.

271 The identical instruction lands Gemini 3.1 Pro, GLM-5.2, and Claude Opus 4.8 at 0.071 , 0.061 , and
 272 -0.037 , two significantly short of the human articles and one significantly past them. The wording
 273 alone moves Claude Opus 4.8 from significantly short of the human articles in control and few-shot
 274 to significantly past them under instruction and self-refine.

275 All nine families undershoot significantly under few-shot. The few-shot gap exceeds the family’s
 276 control gap in six of nine, and ranked by rank-biserial it is the largest-effect mode in seven of nine
 277 (0.94 on Gemini 3.1 Pro and GLM-5.2, 0.83 on Kimi K2.6, 0.80 on Qwen3.6-27B, 0.76 on Grok
 278 4.3, 0.69 on Grok 4.5, 0.65 on Gemma 3 27B). The exceptions, DeepSeek V4 Pro and Claude Opus
 279 4.8, are the families whose uninstructed control carries the larger effect. On Grok 4.5 few-shot alone
 280 drags the model back below the human level (0.077 , rank-biserial 0.69 , $p < 10^{-5}$). The exemplars
 281 are three human news reports at the median of the human rhythm distribution (Appendix A), so
 282 showing the model typical human rhythm produces text farther from it than saying nothing, in six of
 283 nine families; one reading we did not test is register reproduction rather than variance.

284 The synthesis is the scatter. Across the forty-five family-by-mode cells, thirty-two land significantly
 285 flatter than the human articles, eight show no significant difference, and five land significantly more
 286 dispersed. The same instruction, word for word, leaves four families significantly short of the human
 287 dispersion level, lands two on it, moves one further away, and pushes two significantly past. Prompt-
 288 level control has poor aim: its gain and even its sign are family-dependent, and the uninstructed
 289 gap does not predict the landing. This is what the diagnostic exposes: the gap it measures cannot
 290 be closed reliably from the prompt side, because the instruction that reaches parity on one family
 291 moves the next one a different amount and sometimes the wrong way.

292 5 Discussion

293 What this audit prescribes for evaluation practice.

- 294 1. Normalize before scoring: any segmenter-based feature computed on raw model output
295 measures formatting as well as rhythm, and the unreported choice can flip verdicts in both
296 directions (Section 4.2).
- 297 2. Sample multiple seeds and average within headline before testing: pooling seeds pseudo-
298 replicates the human baseline and overstates significance (Section 4.3).
- 299 3. Audit empty cells: silent provider moderation returns ordinary responses with empty output,
300 and unlogged empties reshape scores invisibly (Section 4.4).
- 301 4. State parity as CI-bounded equivalence against pre-specified margins and report the small-
302 est margin cleared; a failure to reject is not a match (Section 4.3).

303 **Trusting burstiness as a detection signal.** Because burstiness is the feature detectors lean on, a
304 definition that raises it could be read as an anti-detection tool. We resist that framing on scientific
305 and ethical grounds. For generation, the plain instruction’s landing is family-dependent, significantly
306 short of the human dispersion level on five of nine families, on it on two, and past it on two. For de-
307 tection, the instructed dispersion signature persists in five of nine families, but Grok 4.5, Kimi K2.6,
308 and Claude Opus 4.8 each write at or beyond human dispersion under at least one prompted mode,
309 where a flatness threshold fires in the wrong direction, and any segmenter-based feature computed
310 on raw output inherits the formatting artifacts of Section 4.2. We report detector-relevant behavior
311 across the construct rather than optimizing to defeat any one detector, and the strong detector signal
312 (surprisal) stays outside it.

313 **Dual use and disclosure.** We disclose AI-assisted research per the venue policy, and we bench-
314 mark against detectors transparently.

315 **What the diagnostic leaves open.** For the five modes tested, the landing cannot be aimed: the
316 same wording lands families short of, on, and past the human level, and the uninstructed gap does
317 not predict which.

318 6 Limitations

319 **Limits of the metric.** Statistical separation is not construct validity: significant paired differences
320 show the coefficient separates authorship in matched news prose, and they leave open whether it
321 captures everything meant by rhythm. The coefficient B_n inherits the weaknesses of the coefficient
322 of variation it reparametrizes (Section 3): it is order-blind and shape-blind, scoring the dispersion of
323 the length sequence while ignoring where the long and short sentences fall, and memory recovers
324 only lag-1 order. Surprisal is model-relative, which is why it stays outside the construct and is not
325 scored. The metric is unreliable below its ten-sentence floor (Section 3) and is genre-relative: target
326 bands do not transfer unchanged across registers. Syntactic complexity, a second intended axis of
327 structural variation, is work in progress and not part of the locked definition. Guard thresholds are
328 heuristic: they set a floor and prove nothing about strategies outside their list.

329 **Limits of the experiment.** The corpus is a single register (Wikinews, expository news, English),
330 chosen as a conservative, low-rhythm register, so reported gaps are likely smaller than on narrative
331 or literary text. Nine families occupy a narrow band of contemporary instruction-tuned assistants,
332 the Grok pair contributing two successive generations; for budget reasons we omitted an OpenAI
333 leg. Results are not yet shown across genres, languages, or future vintages, and burstiness norms are
334 themselves genre- and language-relative. The flatness tell is model- and vintage-relative: Grok 4.3
335 to Grok 4.5 flips the instruction-mode gap in a single generation, so the gaps we measure should not
336 be read as constants of nature. One fixed wording per mode risks measuring one model’s reading of
337 one prompt. Cross-family replication removes part of that risk, since few-shot is the largest-effect
338 mode by rank-biserial in seven of nine families (DeepSeek V4 Pro and Claude Opus 4.8 are the
339 exceptions). DeepSeek V4 Pro’s three metric-free paraphrases land instruction at or past the human
340 articles on B_n (median deltas -0.026 to -0.030 against the canonical wording’s $+0.000$), so the
341 mode conclusion holds under rewording, with a wording-sensitivity band of about 0.03.

342 Decoding is not matched across providers (four legs do not honor a temperature setting, two ran
343 at provider defaults, one closed leg ran at temperature 1.0), so cross-family magnitude compar-
344 isons carry a decoding confound while within-family comparisons do not, since every mode of a

leg ran under the same decoding. Each twenty-cell grid receives a Benjamini–Hochberg correction (Appendix B), and no correction spans families. Of the five past-human cells, four survive both corrections; Grok 4.5’s instruction inversion is exploratory under Holm. Where we claim equivalence we bound it, but the margins are analysis-time choices: 0.03 on a $[-1, 1]$ coefficient is a judgment about practical insignificance.

The human articles may appear in the generators’ training data: the 2016–2021 corpus predates all nine training cutoffs; we measured overlap directly. An n -gram probe over 600 control-mode generations per model on three families finds mean matched 8-gram overlap with the paired article of 0.00055 (Qwen3.6-27B, max 0.036), 0.00207 (DeepSeek V4 Pro, max 0.104), and 0.00098 (Grok 4.3, max 0.062), against a mismatched-article baseline mean of 0.0 everywhere; 13-gram means are at or below 0.00064 for all three. The worst single generation shares roughly 10% of its 8-grams with its paired article, consistent with a shared quotation rather than memorization. Verbatim reproduction therefore cannot explain the measured gaps on the probed families. Memorization would move dispersion toward the human level, so any unprobed effect, whether a deficit or an overshoot, is, if contaminated, an underestimate. The direction reverses for the parity landings: memorization would inflate instructed dispersion to the human level, so contamination could manufacture the equivalences we report. Both fall in the instruction mode the probe did not cover; DeepSeek V4 Pro’s control-mode overlap (0.00207) bounds it only indirectly, and the probe did not cover Kimi K2.6 at all.

7 Conclusion

Burstiness has been measured everywhere and, in detection practice, formalized only in an unreviewed detector note. This paper gives it a reproducible, model-free definition hardened against gaming, and surprisal stays outside. Its first use is the aim problem: on uniformly normalized text, eight of nine families write significantly flatter than the human articles uninstructed. The same instruction leaves four families significantly short of the human dispersion level, lands two on it, moves one further away, and pushes two significantly past it. Few-shot exemplars are the flattest mode by rank-biserial in seven of nine.

The contribution is the composition: a principled, model-free definition, guarded against gaming, scored on uniformly normalized text, and exercised in a matched paired design across nine families. The gap it exposes is one that prompting moves but cannot reliably close.

References

- Advik Raj Basani and Pin-Yu Chen. Diversity Boosts AI-Generated Text Detection. arXiv, 2025. URL <https://arxiv.org/abs/2509.18880>.
- Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language Models are Few-Shot Learners. NeurIPS 2020, 2020. URL <https://arxiv.org/abs/2005.14165>.
- Kathleen C. Fraser, Hillary Dawkins, and Svetlana Kiritchenko. Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods. arXiv, 2024. URL <https://arxiv.org/abs/2406.15583>.
- Corrado Gini. *Variabilità e Mutabilità*. C. Cuppini, Bologna, 1912. URL <https://ui.adsabs.harvard.edu/abs/1912vamu.book.....G/abstract>.
- Kwang-Il Goh and Albert-László Barabási. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 2008. doi: 10.1209/0295-5075/81/48002. URL <https://doi.org/10.1209/0295-5075/81/48002>.
- Eun-Kyeong Kim and Hang-Hyun Jo. Measuring burstiness for finite event sequences. *Physical Review E*, 2016. doi: 10.1103/PhysRevE.94.032311. URL <https://doi.org/10.1103/PhysRevE.94.032311>.
- Aman Madaan, Niket Tandon, Prakhar Gupta, et al. Self-Refine: Iterative Refinement with Self-Feedback. NeurIPS 2023, 2023. URL <https://arxiv.org/abs/2303.17651>.

393 Jií Milíka, Anna Marklová, and Václav Cvrlek. Benchmark of Stylistic Variation in LLM-Generated
394 Texts. arXiv, 2025. URL <https://arxiv.org/abs/2509.10179>.

395 Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke
396 Zettlemoyer. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?
397 EMNLP 2022, 2022. URL <https://arxiv.org/abs/2202.12837>.

398 Tara Radvand, Mojtaba Abdolmaleki, Mohamed Mostagir, and Ambuj Tewari. A Training-free
399 Method for LLM Text Attribution. arXiv, 2025. URL <https://arxiv.org/abs/2501.02406>.

400 Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani.
401 Evaluating the Diversity and Quality of LLM Generated Content. arXiv, 2025. URL <https://arxiv.org/abs/2504.12522>.

403 smail Tarm and Aytu Onan. Can You Detect the Difference? arXiv, 2025. URL <https://arxiv.org/abs/2507.10475>.

405 Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim.
406 Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Lan-
407 guage Models. ACL 2023, 2023. URL <https://arxiv.org/abs/2305.04091>.

408 Wikinews. Wikinews: free-content news source (English, 2016-2021), 2021. URL <https://en.wikinews.org>.

410 C. B. Williams. A note on the statistical analysis of sentence-length as a criterion of literary style.
411 *Biometrika*, 1940. URL [https://academic.oup.com/biomet/article-pdf/31/3-4/356/](https://academic.oup.com/biomet/article-pdf/31/3-4/356/499025/31-3-4-356.pdf)
412 [499025/31-3-4-356.pdf](https://academic.oup.com/biomet/article-pdf/31/3-4/356/499025/31-3-4-356.pdf).

413 G. Udny Yule. On sentence-length as a statistical characteristic of style in prose. *Biometrika*,
414 1939. URL [https://pure.mpg.de/pubman/item/item_2407783_2/component/file_](https://pure.mpg.de/pubman/item/item_2407783_2/component/file_2407782/Yule_1939_On_Sentence_Lenght_Biometrika_1939.pdf)
415 [2407782/Yule_1939_On_Sentence_Lenght_Biometrika_1939.pdf](https://pure.mpg.de/pubman/item/item_2407783_2/component/file_2407782/Yule_1939_On_Sentence_Lenght_Biometrika_1939.pdf).

416 A Prompt wordings

417 The five prompting modes share one length-matched task brief and, apart from the control mode, one
418 persona frame. This appendix gives the exact wording of every component, the rule that composes
419 them into each mode, and two rules the wording leaves implicit: how the target length is set and how
420 the few-shot exemplars are chosen. The byte-exact source is `bench/prompts.py` in the released
421 code, which also carries a self-check that enforces the three-call structure of the self-refine mode
422 and the metric-free property stated below. In the blocks that follow, lines beginning with `#` are
423 annotations; every other line is the literal prompt string, with the long single-line strings wrapped
424 to fit the page and blank lines marking the paragraph breaks inside a string. Braced tokens such as
425 `{headline}` are filled per generation.

```
426 # BRIEF: control mode sends only this
427 Write a news article with the following headline:
428
429 {headline}
430
431 Aim for about {target_words} words.
432
433 # VOICE: prepended to every non-control mode
434 Act as a professional news journalist. Provide the article a seasoned
435 staff reporter would write: concrete and specific, in a natural human
436 voice, never flat or mechanical.
437
438 # TAIL: appended to every non-control mode
439 Keep that natural, human quality all the way through, including the
440 closing paragraphs.
```

```

442 # PLAN: plan-mode body, between VOICE and BRIEF
443 Before you write, plan the article the way a reporter does: pick the
444 lede (the single most important fact), then lay out the who, what,
445 when, where, why, and how, and the order to present them, most
446 important first. Work that outline out for yourself, then write the
447 article to it. Do not show your planning; return only the finished
448 article.
449
450 # FEWSHOT_LEAD / FEWSHOT_TAIL: wrap the exemplar block
451 Study the voice, pacing, and rhythm of the examples below, then write
452 in the same spirit. Borrow only their style, never their facts or
453 wording.
454
455 <examples>
456 <example>
457 # exemplar 1 (full article text omitted; see "Few-shot exemplars" below)
458 </example>
459 # ... two more <example> blocks, three exemplars in total ...
460
461 </examples>
462
463 # FEEDBACK: self-refine call 2, over call 1's {draft}
464 You are a demanding editor. Critique the draft below only on how
465 naturally human its prose reads: its rhythm, pacing, and flow. Name
466 at least THREE specific passages that read as flat, monotonous, or
467 mechanical. For each, quote the exact sentence(s) and say concretely
468 what to change to give it a more human cadence. Do not rewrite the
469 draft; give only the critique. Even if the writing is already strong,
470 still name the three weakest spots.
471
472 DRAFT:
473 {draft}
474
475 # REWRITE: self-refine call 3, over {draft} and {feedback}
476 Here is your draft and an editor's critique.
477
478 DRAFT:
479 {draft}
480
481 CRITIQUE:
482 {feedback}
483
484 Rewrite the article in full, applying every point of the critique. This
485 is a substantial rewrite, not a light edit: change how the sentences
486 move wherever the critique flagged flatness, while preserving the
487 meaning, the facts, and the target length. Return only the rewritten
488 article.

```

489 **How the modes compose.** The control mode sends the bare brief and nothing else. Every other
490 mode follows one template: the voice preamble, then a mode-specific body, then the same brief, then
491 the tail. The plain instruction has an empty body, so it reduces to the voice preamble, the brief, and
492 the tail. The plan mode puts the plan text in the body slot, and the few-shot mode puts the exemplar
493 block there. Self-refine is three model calls on a single article: a draft from the instruction prompt,
494 a critique from the feedback prompt applied to that draft, and a rewrite from the rewrite prompt
495 applied to the draft and the critique. A self-check in `bench/prompts.py` asserts the three calls.

496 **Length brief.** The `{target_words}` slot is the paired human article's word count rounded to the
497 nearest 25, so every generation targets its human counterpart's length (Section 4.1).

498 **Few-shot exemplars.** The exemplar block holds three straight news reports drawn from Wikinews
 499 articles outside the 200 evaluation set. They are the reports whose burstiness sits nearest the median
 500 of the human distribution, chosen for register and topic. The selection uses B_n only to locate that
 501 median and drop degenerate tails, never to pick high-dispersion exemplars, so the few-shot condition
 502 sees typical human rhythm and no exemplar is chosen to teach to the test. Their full texts are
 503 committed with the released code and are omitted here for length.

504 **Metric-free wording.** None of the prompts names the axis the benchmark scores. A self-check
 505 enforces that no prompt string contains a measured-axis term, among them *burstiness*, *sentence*
 506 *length*, *variance*, and *standard deviation*. No prompt gives the model a numeric target on a measured
 507 axis; the word count in the length brief is a length target only. Keeping the axis out of the wording
 508 guards against teaching to the test.

509 **Instruction paraphrases (wording-robustness probe).** The probe in Section 6 re-runs the in-
 510 struction mode on DeepSeek V4 Pro under three independent rewrites of the voice and tail above,
 511 one seed over all 200 briefs, each keeping the persona frame and the metric-free property. The
 512 harness is experiments/ablation/prompt_paraphrase.py. The three pairs follow, the voice
 513 prepended and the tail appended around the same brief.

```
514 # PARAPHRASE v1 (VOICE then TAIL)
515 You are a veteran newspaper reporter. Write the story the way
516 an experienced newsroom professional would: specific, grounded
517 in facts, with a voice that reads as unmistakably human rather
518 than machine-written.
519
520 Hold that authentic reporter’s voice through to the final sentence.
521
522 # PARAPHRASE v2 (VOICE then TAIL)
523 Take the role of a seasoned journalist at a major daily. Produce
524 the kind of article an editor would run with light touches only:
525 concrete details, a living, natural voice, nothing stiff or
526 formulaic.
527
528 Make sure the writing still sounds fully human in the closing
529 paragraphs, not just the opening.
530
531 # PARAPHRASE v3 (VOICE then TAIL)
532 Write as a long-time staff correspondent. The article should
533 carry the texture of real human reportage: precise facts,
534 natural phrasing, and none of the flat, mechanical quality of
535 generated text.
536
537 Sustain that human texture to the very end of the piece.
```

538 B Per-family result grids

539 Every family ran the identical full design: the same 200 matched headlines, the same five mode
 540 wordings (Appendix A), three seeds, uniform normalization, and the same paired scoring on all
 541 four dimensions. Qwen3.6-27B’s grid is Table 2, and the eight grids that follow complete the set,
 542 generated from each family’s released evaluation with no hand-transcribed values. Within each
 543 family we apply Benjamini–Hochberg at $q = 0.05$ across its twenty cells and report the census in
 544 the caption, with the stricter Holm count alongside, and no correction is applied across families.
 545 Instruction-mode equivalence margins are released for every family with the evaluation artifacts.
 546 The two families whose instruction-mode B_n CI clears a pre-specified margin are DeepSeek V4
 547 Pro (0.03) and Kimi K2.6 (0.05) (Section 4.4); Kimi’s instructed leg shows no per-cell significant
 548 difference on any dimension, though only its dispersion cell carries an equivalence bound. Memory
 549 M is where three few-shot legs (Qwen3.6-27B, Kimi K2.6, Claude Opus 4.8) overshoot the human
 550 articles even as their dispersion stays significantly flat. Seg. entropy, the axis most exposed to

Table 2: Qwen3.6-27B, full grid (uniformly normalized scoring): seed-averaged Wilcoxon signed-rank paired tests over 200 matched headlines per mode. Reading key: median Δ = human – model; positive means the model is *below* the human article on that dimension, the flatness direction. Matched rank-biserial in parentheses; * $p < 0.05$, ** $p < 0.01$ (per-cell, two-sided). Within-family census: 13 of 20 cells per-cell significant; 12 survive both Benjamini–Hochberg at $q = 0.05$ and Holm (the exception, instruction-mode memory, is exploratory; Section 4.3).

Mode	B_n	Gini	Seg. entropy	Memory M
Control	0.097 (0.67)**	0.046 (0.67)**	−0.001 (0.10)	−0.007 (−0.05)
Instruction	0.028 (0.34)**	0.010 (0.28)**	−0.018 (−0.04)	−0.052 (−0.16)*
Few-shot	0.119 (0.80)**	0.060 (0.79)**	0.127 (0.47)**	−0.071 (−0.26)**
Plan	0.045 (0.41)**	0.017 (0.35)**	0.057 (0.16)	−0.028 (−0.07)
Self-refine	0.039 (0.26)**	0.021 (0.24)**	0.029 (0.10)	0.028 (0.09)

551 formatting artifacts, still separates on normalized text: instructed DeepSeek V4 Pro and four of
552 Grok 4.5’s five modes sit significantly above the human articles.

553 Table 2 reports Qwen3.6-27B’s paired deltas. The uninstructed control differs from the human
554 articles where dispersion is measured: median $\Delta B_n = 0.097$ (rank-biserial 0.67, $p < 10^{-5}$), with
555 Gini concurring (0.046, rank-biserial 0.67). On segmentation entropy (−0.001, $p = 0.24$) and
556 memory (−0.007, $p = 0.52$) the control mode shows no significant difference. The control-mode
557 authorship gap is confined to the dispersion dimensions.

558 A plain instruction narrows the dispersion gap by roughly seven tenths and does not close it. ΔB_n
559 falls from 0.097 to 0.028 (rank-biserial 0.34, $p = 3 \times 10^{-5}$) and Gini from 0.046 to 0.010
560 ($p < 10^{-3}$); both effects survive correction. Against the equivalence margins, instruction-mode B_n
561 is not equivalent at 0.03 (CI [0.013, 0.046], $m^* = 0.046$) and becomes equivalent only at 0.05. Gini
562 is equivalent at 0.03 ($m^* = 0.024$), segmentation entropy at 0.08 ($m^* = 0.063$, CI [−0.063, 0.045]),
563 and memory at no margin we test ($m^* = 0.095$, CI [−0.095, −0.003]), though its per-cell signifi-
564 cance is the one exploratory cell of Section 4.3.

565 Few-shot exemplars are the flattest mode on the dispersion statistics, a B_n deficit of 0.119 with
566 rank-biserial 0.80, and the per-seed effect holds at 0.724–0.745 on every individual seed.

567 Plan and self-refine behave like weaker instructions: significant dispersion deficits (B_n 0.045 and
568 0.039, rank-biserial 0.41 and 0.26, both $p < 0.01$) with no correction-surviving effect on segmen-
569 tation entropy (plan 0.057, $p = 0.052$; self-refine 0.029, $p = 0.22$) or memory. All ten dispersion
570 cells (five modes by B_n and Gini) are significant at $p < 0.01$ and survive correction, and the only
571 other survivors are few-shot’s segmentation entropy (0.127) and memory (−0.071) cells.

Table 3: Gemini 3.1 Pro, full grid: seed-averaged Wilcoxon paired tests over 200 matched headlines per mode; median Δ = human – model, matched rank-biserial in parentheses; * $p < 0.05$, ** $p < 0.01$ per cell. Within-family census: 16 of 20 cells per-cell significant; all 16 survive Benjamini–Hochberg at $q = 0.05$ (13 survive Holm).

Mode	B_n	Gini	Seg. entropy	Memory M
Control	0.128 (0.85)**	0.056 (0.81)**	−0.062 (−0.19)*	0.043 (0.18)*
Instruction	0.071 (0.56)**	0.026 (0.47)**	−0.053 (−0.16)	0.109 (0.38)**
Few-shot	0.169 (0.94)**	0.080 (0.93)**	0.117 (0.39)**	0.051 (0.24)**
Plan	0.070 (0.58)**	0.028 (0.48)**	−0.026 (−0.06)	0.114 (0.47)**
Self-refine	0.039 (0.27)**	0.013 (0.21)**	−0.051 (−0.12)	0.047 (0.12)

572 C Reference implementation

573 The measurement instrument is model-free and pure standard library. The listing below is the com-
574 plete, self-contained implementation used in this paper: the sentence segmenter, the burstiness co-
575 efficient B_n (Goh–Barabási under the Kim–Jo finite- n correction), the memory coefficient M , the
576 Gini robustness check, Miller–Madow segmentation entropy, the multiscale profile, and the Good-

Table 4: Gemma 3 27B, full grid (design and conventions as in Table 2: median Δ = human – model, matched rank-biserial in parentheses; * $p < 0.05$, ** $p < 0.01$ per cell). Within-family census: 16 of 20 cells per-cell significant; all 16 survive both Benjamini–Hochberg at $q = 0.05$ and Holm.

Mode	B_n	Gini	Seg. entropy	Memory M
Control	0.069 (0.58)**	0.037 (0.58)**	0.063 (0.25)**	0.023 (0.11)
Instruction	0.049 (0.44)**	0.025 (0.41)**	0.086 (0.36)**	0.040 (0.10)
Few-shot	0.063 (0.65)**	0.033 (0.63)**	0.145 (0.48)**	0.005 (0.02)
Plan	0.061 (0.57)**	0.033 (0.55)**	0.129 (0.45)**	0.003 (0.03)
Self-refine	0.029 (0.30)**	0.015 (0.29)**	0.094 (0.32)**	0.048 (0.21)**

Table 5: DeepSeek V4 Pro, full grid (design and conventions as in Table 2: median Δ = human – model, matched rank-biserial in parentheses; * $p < 0.05$, ** $p < 0.01$ per cell). Within-family census: 8 of 20 cells per-cell significant; all 8 survive Benjamini–Hochberg at $q = 0.05$ (6 survive Holm). The instruction-mode B_n cell is the CI-bounded equivalence result of Section 4.4.

Mode	B_n	Gini	Seg. entropy	Memory M
Control	0.035 (0.34)**	0.018 (0.35)**	−0.050 (−0.12)	−0.025 (−0.07)
Instruction	0.000 (0.05)	−0.005 (−0.02)	−0.145 (−0.47)**	0.006 (0.07)
Few-shot	0.029 (0.24)**	0.014 (0.20)*	−0.068 (−0.26)**	0.012 (0.01)
Plan	0.013 (0.15)	0.003 (0.08)	−0.117 (−0.36)**	−0.003 (0.03)
Self-refine	−0.018 (−0.10)	−0.010 (−0.11)	−0.139 (−0.49)**	0.027 (0.14)

Table 6: Grok 4.3, full grid (design and conventions as in Table 2: median Δ = human – model, matched rank-biserial in parentheses; * $p < 0.05$, ** $p < 0.01$ per cell). Within-family census: 13 of 20 cells per-cell significant; all 13 survive Benjamini–Hochberg at $q = 0.05$ (9 survive Holm).

Mode	B_n	Gini	Seg. entropy	Memory M
Control	0.027 (0.27)**	0.010 (0.23)**	−0.046 (−0.16)	−0.004 (0.01)
Instruction	0.050 (0.42)**	0.026 (0.39)**	0.062 (0.19)*	0.028 (0.10)
Few-shot	0.108 (0.76)**	0.052 (0.74)**	0.180 (0.56)**	0.001 (−0.05)
Plan	0.061 (0.56)**	0.026 (0.50)**	0.136 (0.35)**	0.023 (0.03)
Self-refine	0.019 (0.19)*	0.009 (0.14)	0.055 (0.21)**	0.028 (0.14)

Table 7: Grok 4.5, full grid (design and conventions as in Table 2: median Δ = human – model, matched rank-biserial in parentheses; * $p < 0.05$, ** $p < 0.01$ per cell). Negative cells place the model above the human articles, more dispersed than the human baseline. Within-family census: 14 of 20 cells per-cell significant; 13 survive Benjamini–Hochberg at $q = 0.05$ (control-mode Gini, $p = 0.037$, is the loss; 9 survive Holm).

Mode	B_n	Gini	Seg. entropy	Memory M
Control	−0.009 (−0.14)	−0.011 (−0.17)*	−0.249 (−0.74)**	−0.023 (−0.07)
Instruction	−0.025 (−0.20)*	−0.014 (−0.20)*	−0.099 (−0.40)**	0.048 (0.22)**
Few-shot	0.077 (0.69)**	0.039 (0.68)**	0.089 (0.28)**	−0.002 (−0.05)
Plan	−0.009 (−0.02)	−0.005 (−0.05)	−0.062 (−0.22)**	0.021 (0.06)
Self-refine	−0.059 (−0.41)**	−0.033 (−0.44)**	−0.133 (−0.44)**	0.071 (0.31)**

Table 8: GLM-5.2, full grid (design and conventions as in Table 2: median Δ = human – model, matched rank-biserial in parentheses; * $p < 0.05$, ** $p < 0.01$ per cell). Within-family census: 16 of 20 cells per-cell significant; 16 survive Benjamini–Hochberg at $q = 0.05$ (14 survive Holm). 36 silent-refusal empty cells are excluded by the seed-averaging protocol (Section 4.4).

Mode	B_n	Gini	Seg. entropy	Memory M
Control	0.126 (0.85)**	0.058 (0.82)**	0.068 (0.25)**	0.053 (0.19)*
Instruction	0.061 (0.57)**	0.029 (0.51)**	0.064 (0.26)**	0.076 (0.24)**
Few-shot	0.148 (0.94)**	0.068 (0.92)**	0.221 (0.74)**	0.022 (0.07)
Plan	0.080 (0.65)**	0.034 (0.59)**	0.091 (0.32)**	0.034 (0.20)*
Self-refine	0.011 (0.13)	0.002 (0.09)	0.050 (0.33)**	−0.019 (−0.02)

Table 9: Kimi K2.6, full grid (design and conventions as in Table 2: median Δ = human – model, matched rank-biserial in parentheses; * $p < 0.05$, ** $p < 0.01$ per cell). Within-family census: 12 of 20 cells per-cell significant; 10 survive Benjamini–Hochberg at $q = 0.05$ (9 survive Holm). 25 provider content-filter refusals are excluded by the seed-averaging protocol (Section 4.4).

Mode	B_n	Gini	Seg. entropy	Memory M
Control	0.095 (0.69)**	0.043 (0.63)**	−0.048 (−0.16)*	−0.025 (−0.09)
Instruction	−0.006 (−0.00)	−0.014 (−0.14)	−0.024 (−0.11)	−0.014 (0.08)
Few-shot	0.123 (0.83)**	0.063 (0.79)**	0.152 (0.55)**	−0.073 (−0.26)**
Plan	0.026 (0.24)**	0.006 (0.09)	0.032 (0.05)	0.055 (0.17)*
Self-refine	−0.070 (−0.57)**	−0.040 (−0.58)**	0.061 (0.16)	−0.052 (−0.23)**

Table 10: Claude Opus 4.8, full grid (design and conventions as in Table 2: median Δ = human – model, matched rank-biserial in parentheses; * $p < 0.05$, ** $p < 0.01$ per cell). Within-family census: 16 of 20 cells per-cell significant; 16 survive Benjamini–Hochberg at $q = 0.05$ (15 survive Holm). Ran at effort high with a 32,000-token cap; both deviations and their measurements are recorded in the leg metadata (Section 4.4).

Mode	B_n	Gini	Seg. entropy	Memory M
Control	0.119 (0.80)**	0.054 (0.77)**	0.116 (0.39)**	−0.069 (−0.24)**
Instruction	−0.037 (−0.25)**	−0.031 (−0.36)**	0.032 (0.09)	−0.117 (−0.38)**
Few-shot	0.039 (0.34)**	0.012 (0.26)**	0.253 (0.75)**	−0.045 (−0.25)**
Plan	−0.000 (0.01)	−0.004 (−0.10)	0.108 (0.34)**	−0.059 (−0.25)**
Self-refine	−0.065 (−0.50)**	−0.043 (−0.59)**	−0.005 (0.01)	0.039 (0.18)*

hart guard, whose thresholds are its default keyword arguments. It runs on any Python 3 interpreter with no third-party dependencies and no reference model.

```

579 """Reference implementation of the burstiness measurement instrument.
580
581 Model-free and pure standard library. Segments a document into sentence
582 lengths and computes the burstiness coefficient  $B_n$  (Goh-Barabasi under the
583 Kim-Jo finite- $n$  correction), the memory coefficient  $M$ , the Gini robustness
584 check, Miller-Madow segmentation entropy, and a multiscale profile. The
585 Goodhart guard screens degenerate text that would game a dispersion score.
586 """
587 from __future__ import annotations
588
589 import math
590 import re
591 from statistics import fmean, pstdev
592 from typing import List, Optional, Sequence
593
594 # Below this many sentences,  $B_n$  and  $M$  estimates are noisy and biased; the
595 # entry point reports "reliable = False" rather than trusting them.
596 RELIABILITY_FLOOR = 10
597
598 # -----
599 # Segmentation: sentence lengths in whitespace tokens (stdlib only)
600 # -----
601
602 _ABBREVIATIONS = {
603     "mr", "mrs", "ms", "dr", "prof", "sr", "jr", "st", "vs", "etc",
604     "e.g", "i.e", "fig", "no", "vol", "cf", "al", "inc", "ltd", "co",
605 }
606 _SENT_SPLIT = re.compile(r'([.!?]+["\']\s*)\s+')
607 _PUNCT_SYMBOLS = re.compile(r'[.!?,:;-()"\']')
608
609
610 def sentence_lengths(text: str) -> List[int]:
611     """Word-token count of each sentence, with an abbreviation-guarded split."""
612     if not text or not text.strip():
613         return []

```

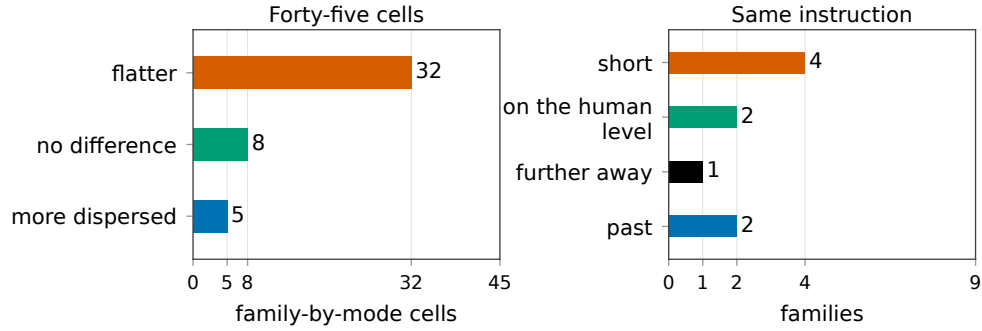


Figure 2: Left: of the forty-five family-by-mode cells, thirty-two land significantly flatter than the human articles, eight show no significant difference, and five land significantly more dispersed. Right: the same instruction leaves four families significantly short of the human dispersion level, lands two on it, moves one further away, and pushes two significantly past (Section 4.4).

```

614     sentences: List[str] = []
615     last = 0
616     for m in _SENT_SPLIT.finditer(text):
617         end = m.end()
618         candidate = text[last:end].strip()
619         bare = candidate.rstrip('!.?"/\']').strip()
620         last_word = bare.split()[-1].lower() if bare.split() else ""
621         if last_word in _ABBREVIATIONS:
622             continue
623         if candidate:
624             sentences.append(candidate)
625         last = end
626     tail = text[last:].strip()
627     if tail:
628         sentences.append(tail)
629     return [len(s.split()) for s in sentences if s.split()]
630
631
632     # -----
633     # Core coefficients (operate on a sequence of sentence lengths)
634     # -----
635
636     def burstiness_B(intervals: Sequence[float]) -> Optional[float]:
637         """Goh-Barabasi coefficient B = (sigma - mu) / (sigma + mu), bounded [-1, 1]."""
638         if len(intervals) < 2:
639             return None
640         mu = fmean(intervals)
641         sigma = pstdev(intervals)
642         denom = sigma + mu
643         if denom <= 0:
644             return None
645         return (sigma - mu) / denom
646
647
648     def burstiness_B_corrected(intervals: Sequence[float]) -> Optional[float]:
649         """Kim-Jo finite-n correction B_n; reduces to (r-1)/(r+1) as n -> inf."""
650         n = len(intervals)
651         if n < 2:
652             return None
653         mu = fmean(intervals)
654         if mu <= 0:
655             return None
656         r = pstdev(intervals) / mu
657         sp, sm = math.sqrt(n + 1), math.sqrt(n - 1)
658         denom = (sp - 2.0) * r + sm
659         if denom == 0:

```

```

660         return None
661     return (sp * r - sm) / denom
662
663
664 def memory_M(intervals: Sequence[float]) -> Optional[float]:
665     """Lag-1 autocorrelation of consecutive sentence lengths (the memory term)."""
666     n = len(intervals)
667     if n < 3:
668         return None
669     a = list(intervals[:-1])
670     b = list(intervals[1:])
671     m1, m2 = fmean(a), fmean(b)
672     s1, s2 = pstdev(a), pstdev(b)
673     if s1 <= 0 or s2 <= 0:
674         return None
675     cov = fmean((x - m1) * (y - m2) for x, y in zip(a, b))
676     return cov / (s1 * s2)
677
678
679 def gini(intervals: Sequence[float]) -> float:
680     """Gini coefficient of sentence lengths; scale-free robustness check for B_n."""
681     n = len(intervals)
682     if n < 2:
683         return 0.0
684     m = fmean(intervals)
685     if m == 0:
686         return 0.0
687     # Sample (unbiased) Gini: divide by n*(n-1), not n^2, which removes an
688     # n-dependent bias that would confound cross-document comparison.
689     return sum(abs(a - b) for a in intervals for b in intervals) / (2.0 * n * (n - 1) * m)
690
691
692 _TYPO_QUOTES = str.maketrans({"": "'", "'": "'", "": "'", "": "'"})
693
694
695 def segmentation_entropy(text: str) -> float:
696     """Miller-Madow-corrected Shannon entropy (bits) of the punctuation marks.
697
698     Typographic quotes are normalized to ASCII first, so quote STYLE (curly vs
699     straight) is not scored as segmentation diversity.
700     """
701     counts: dict = {}
702     for ch in _PUNCT_SYMBOLS.findall((text or "").translate(_TYPO_QUOTES)):
703         counts[ch] = counts.get(ch, 0) + 1
704     total = sum(counts.values())
705     if total == 0:
706         return 0.0
707     h = -sum((c / total) * math.log2(c / total) for c in counts.values())
708     return h + (len(counts) - 1) / (2.0 * total * math.log(2))
709
710
711 # -----
712 # Multiscale temporal profile
713 # -----
714
715 def _coarse_grain(L: Sequence[float], w: int) -> List[float]:
716     """Non-overlapping window means: coarsen the length series to scale w."""
717     return [fmean(L[i:i + w]) for i in range(0, len(L) - w + 1, w)]
718
719
720 def multiscale_burstiness(text: str, scales: Sequence[int] = (1, 2, 4)) -> dict:
721     """B_n at the sentence scale and at coarse-grained window scales.
722
723     Genuine long-range burstiness holds its dispersion at coarse scales, whereas
724     mechanical alternation washes out toward B_n = -1 at scale >= 2.
725     """
726     L = [float(x) for x in sentence_lengths(text)]
727     out = {}

```

```

728     for s in scales:
729         cg = list(L) if s == 1 else _coarse_grain(L, s)
730         out[f"B_n@{s}"] = burstiness_B_corrected(cg) if len(cg) >= 2 else None
731     return out
732
733
734 # -----
735 # Goodhart guard: reject degenerate texts that reach the score by gaming it
736 # -----
737
738 def _trigram_diversity(text: str) -> float:
739     """Distinct word-trigram ratio (catches verbatim repetition)."""
740     w = (text or "").split()
741     if len(w) < 4:
742         return 1.0
743     tris = list(zip(w, w[1:], w[2:]))
744     return len(set(tris)) / len(tris)
745
746
747 def _runon_ratio(text: str) -> float:
748     """Internal clause delimiters (, ; :) per sentence terminal (. ! ?)."""
749     internal = len(re.findall(r"[;,:]", text or ""))
750     terminal = len(re.findall(r"[.!?]", text or "")) or 1
751     return internal / terminal
752
753
754 def gaming_flags(text, *, loo_delta=0.25, m_floor=-0.5, m_ceiling=0.7,
755                 distinct_len_min=0.5, trigram_min=0.5, frag_frac_max=0.3,
756                 punct_rate_max=0.25, runon_max=8.0, finescale_drop=0.4):
757     """Flag texts that reach a high dispersion score through degenerate patterns.
758
759     Returns {"ok", "gate_ok", "flags", "detail"}. ‘ok’ means no gaming pattern
760     fired; ‘gate_ok’ means ok AND reliable (n >= RELIABILITY_FLOOR). B_n should
761     be counted only when gate_ok is True and a separate coherence gate also
762     passes: re-punctuating a fixed text can swing B_n across much of its range,
763     so length-rhythm must never select generations on its own.
764
765     Flags: outlier_driven (dropping the longest sentence swings B_n > loo_delta),
766     low_len_diversity (distinct-length ratio < distinct_len_min at n >= 8, which
767     catches blocky and period-k patterns), repetitive (trigram ratio <
768     trigram_min), alternating (M < m_floor), over_clumped (M > m_ceiling),
769     fine_scale_only (B_n@1 > 0 but drops by > finescale_drop at scale 2),
770     fragment_spam (>= frag_frac_max of sentences under 3 words), punct_gamed
771     (punctuation marks per word > punct_rate_max), runon (internal-to-terminal
772     delimiter ratio > runon_max).
773
774     """
775     Li = [int(x) for x in sentence_lengths(text)]
776     L = [float(x) for x in Li]
777     n = len(L)
778     if n < 3:
779         return {"ok": False, "gate_ok": False, "flags": {"too_short": True}, "detail":
780             {"n": n}}
781
782     B_full = burstiness_B_corrected(L)
783     j = max(range(n), key=lambda i: L[i])
784     B_loo = burstiness_B_corrected(L[:j] + L[j + 1:])
785     loo = abs(B_full - B_loo) if (B_full is not None and B_loo is not None) else 0.0
786     M = memory_M(L)
787     distinct_ratio = len(set(Li)) / n
788     trig = _trigram_diversity(text)
789     frag_frac = sum(1 for x in Li if x < 3) / n
790     words = sum(L) or 1.0
791     punct_rate = len(_PUNCT_SYMBOLS.findall(text or "")) / words
792     runon = _runon_ratio(text)
793     ms = multiscale_burstiness(text)
794     b1, b2 = ms.get("B_n@1"), ms.get("B_n@2")
795     fine_scale_only = (b1 is not None and b2 is not None and b1 > 0.0
796         and (b1 - b2) > finescale_drop)

```

```

796
797     flags = {
798         "outlier_driven": loo > loo_delta,
799         "low_len_diversity": n >= 8 and distinct_ratio < distinct_len_min,
800         "repetitive": trig < trigram_min,
801         "alternating": M is not None and M < m_floor,
802         "over_clumped": M is not None and M > m_ceiling,
803         "fine_scale_only": fine_scale_only,
804         "fragment_spam": frag_frac >= frag_frac_max,
805         "punct_gamed": punct_rate > punct_rate_max,
806         "runon": runon > runon_max,
807     }
808     ok = not any(flags.values())
809     return {
810         "ok": ok,
811         "gate_ok": ok and n >= RELIABILITY_FLOOR,
812         "flags": flags,
813         "detail": {"n": n, "loo_dB": round(loo, 3),
814                   "M": None if M is None else round(M, 3),
815                   "distinct_ratio": round(distinct_ratio, 3), "trigram_div": round(trig,
816                   3),
817                   "frag_frac": round(frag_frac, 3), "punct_rate": round(punct_rate, 3),
818                   "runon": round(runon, 2)},
819     }
820
821
822     # -----
823     # Entry point: the full model-free vector for a document
824     # -----
825
826     def burstiness_pp(text: str) -> dict:
827         """Model-free structural burstiness of a document.
828
829         B_n is the primary dispersion coefficient, M the lag-1 memory, Gini the
830         robustness check, and multiscale the temporal profile. Segmentation entropy
831         is a separate axis (see segmentation_entropy) and surprisal is kept outside
832         the construct. ‘reliable’ is False below RELIABILITY_FLOOR sentences.
833         """
834         L = [float(x) for x in sentence_lengths(text)]
835         n = len(L)
836         return {
837             "B_n": burstiness_B_corrected(L),
838             "M": memory_M(L),
839             "gini": gini(L),
840             "multiscale": multiscale_burstiness(text),
841             "n_sentences": n,
842             "reliable": n >= RELIABILITY_FLOOR,
843             "sentence_lengths": [int(x) for x in L],
844         }

```

845 D Protocol and audit detail

846 D.1 Corpus parse mechanics

847 The human side of each pair is the plain article body, regenerable from the source Wikinews wikitext
848 by a deterministic MediaWiki parse (experiments/ablation/clean_wikinews.py) that keeps
849 the body up to the first section heading or reader-comment block, removes templates, drops reference
850 tags with their contents, resolves wikilinks to their display text, and unwraps inline formatting. On
851 a held-out check of 100 articles this parse reproduces the shipped human bodies exactly.

852 D.2 Per-leg completeness

853 Completeness is disclosed per leg. Of Qwen3.6-27B’s 3,000 cells (200 headlines $\times$ 5 modes $\times$
854 3 seeds), 2,996 completed; we excluded the four that exceeded the reasoning budget rather than

855 scoring their incomplete output. We average affected headlines over their remaining seeds, so every
856 mode retains 200 matched pairs. Grok 4.5 reports two empty self-refine outputs, truncations by
857 signature, seed-averaged away. Provider content moderation removes cells on two legs, by two
858 different mechanisms. Moonshot’s filter refuses explicitly: 25 of Kimi K2.6’s 3,000 cells return
859 HTTP 400, one Hong Kong protest headline refused at every mode and seed. Zhipu’s filter refuses
860 silently: 36 of GLM-5.2’s cells return an ordinary HTTP 200 with empty output, concentrated on the
861 same cluster of politically sensitive headlines. The remaining five legs generated all 200 headlines
862 in every mode.

863 **D.3 Per-leg decoding and output caps**

864 GLM-5.2 ignores a requested temperature and Kimi K2.6 rejects one; Grok 4.5 exposes neither tem-
865 perature nor top- p ; Gemini 3.1 Pro ran at temperature 1.0 with provider-managed thinking; Claude
866 Opus 4.8 exposes no temperature. Gemini’s thoughts are never returned, and Gemma 3 has no think-
867 ing mode. Anthropic bills reasoning against the output cap. The shared 16,000-token cap that leaves
868 other families room to write leaves Claude Opus 4.8 with none at maximum effort: the self-refine
869 rewrite spends the whole budget reasoning. We ran that leg at high effort with a 32,000-token cap,
870 which restores a comparable prose budget rather than granting a larger one and matches Gemini 3.1
871 Pro’s thinking level. The leg completed all 3,000 cells with no truncated or empty output.

872 **D.4 Confidence intervals**

873 CIs on median paired deltas are percentile bootstrap intervals: the 2.5th and 97.5th percentiles of
874 the medians of 2,000 resamples of the 200 matched pairs. Per-seed B_n effect sizes are released
875 alongside the evaluation artifacts and are consistent with the seed-averaged results.

876 **D.5 Scaffolding rates**

877 A generation-hygiene guard flagged markdown scaffolding and echoed headline blocks in every
878 family we benchmark, from 23–38% of control-mode generations on one leg (DeepSeek V4 Pro:
879 23% markdown, 38% title lines) to 100% on another (Gemma 3 27B markdown; headline echoes on
880 the Grok 4.5 and Gemini 3.1 Pro legs). All of it is formatting the plain-prose human baseline lacks.

881 **D.6 Contamination probe values**

882 An n -gram contamination probe over 600 control-mode generations per model on three families
883 finds mean matched 8-gram overlap with the paired article of 0.00055 (Qwen3.6-27B, max 0.036),
884 0.00207 (DeepSeek V4 Pro, max 0.104), and 0.00098 (Grok 4.3, max 0.062), against a mismatched-
885 article baseline mean of 0.0 everywhere; 13-gram means are at or below 0.00064 for all three. The
886 worst single generation shares roughly 10% of its 8-grams with its paired article, consistent with
887 a shared quotation rather than memorization. Verbatim reproduction therefore cannot explain the
888 measured gaps on the three probed families, which span open-weights and API-served models. The
889 probe covered three of the nine families, on the control mode only.