FRACTAL SIGNATURE IS A PROPERTY OF THE GENERATOR, NOT THE DOCUMENT: SPLIT-HALF RELIABILITY BOUNDS FOR SCALING-EXPONENT AI-TEXT DETECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling exponents that summarise the fractal structure of language, including the Hurst and Hölder exponents, separate human from machine-generated text at the corpus level. A growing literature nevertheless uses them as per-document detector features. Because each exponent is defined by an asymptotic limit, a per-document value estimated from roughly 400 tokens is necessarily finite-sample. No reliability coefficient has been reported for such an estimate at document length. We decompose the per-document estimate into a population term, a document term, and estimator error, and measure the ratio of the latter two with a pre-registered interleaved split-half check on $223{,}834$ GAGLE documents scored under Gemma-2B and Mistral-7B. The decomposition yields four falsifiable predictions. We test them with a fixed gradient-boosted detector, a within-population classification task, an edit battery that separates token-order from content perturbations, and zero-shot transfer to RAID. At the population level, all 20 cells (domain, generator, exponent) are separated with non-overlapping 95% intervals. At the document level, the Hölder estimate reaches $\rho = 0.063$ and 0.050 under the two scorers, the Hurst estimate at most 0.576, and none of the twenty candidate quantities clears the bar. All four predictions hold. Deleting the 54 fractal features from an 85-feature detector leaves AUROC unchanged to four decimals; the family receives 5.52% of split gain but 0.47% of permutation importance; within machine text, where population factors are balanced by design, it reaches macro-F1 0.198 against 0.582 for all features (chance 0.125); and its $+0.0240$ AUROC gain on RAID is recovered (149%) by replacing each document's fractal features with their group (generator, domain) mean. Under token-order edits, deleting the family raises recall by 1.96 and 8.76 points. Independently, a curvature detector's apparent robustness to paraphrase is an artifact of scorer-attacker lineage: its recall of 0.63–0.86 falls to 1.2–4.5% under attackers outside that lineage. The fractal signature is a property of the generating process, not of the document. We therefore propose an admissibility condition: use a scaling statistic as a document-level representation only after its split-half reliability at the deployment length clears a declared bar. The check requires one split and one correlation and does not depend on the downstream model.

1 INTRODUCTION

Language has a fractal information structure: its per-token complexity is self-similar across scales and exhibits long-range dependence, with a Hurst parameter near 0.7 (Alabdulmohsin et al., 2024). Alabdulmohsin & Steiner (2025) report that the Hölder and Hurst exponents of natural language occupy a narrow band while those of language-model output vary widely, and conclude that "fractal parameters might prove helpful in detecting LLM-generated text," while declining to make the detection claim themselves, because the exponents describe processes rather than documents. A scaling exponent is defined in the limit $R(n)/S(n) \sim n^H$ as $n \to \infty$, whereas every document is finite. Whether a finite-sample estimate resolves anything about the stretch of text itself, rather than

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

only about the process that produced it, depends on the amount of text available, the *support*. No reliability coefficient has been reported for a scaling estimator at document length (Appendix B).

A thermometer with a five-degree error bar can resolve a real half-degree difference between two towns' climates when enough daily readings are averaged. However, a single day's reading cannot identify the town if the instrument's noise exceeds the difference of interest. In the analogy, the fractal exponent is the thermometer, a (domain, generator) population is the town, and a document's token budget is the single day. We test the resulting claim: the exponent has a real population-level signal, but at the token budget of an individual document, it is dominated by estimation noise rather than by document-specific variation. Section 3.3 translates the analogy into falsifiable predictions, and every result in §4 tests one of them.

The step under examination is one that this literature actually takes. Tulchinskii et al. (2023), the closest neighbor of this work, moves explicitly from a distribution-level separation to a per-document score, and that construction is now widespread both as a detector (Guo et al., 2026; West et al., 2025; Liu et al., 2026; Mansouri et al., 2026; Teja et al., 2025; La Cava & Tagarelli, 2026) and as a characterization of the generated text (Du & Tanaka-Ishii, 2025; Yang et al., 2026). What we report is also not the foreseeable imprecision of a power-law fit over 1.6 decades, which would degrade both exponents alike; instead, the two diverge by nearly an order of magnitude in how well each reproduces itself on the same document, 0.576 against 0.063 (primary sources, Appendix C).

We write the per-document estimate as $\widehat{H}_d = \mu(c_d, g_d, T_d) + \delta_d + \eta_d$, a population term, a document term, and estimator error, and measure the two sides of it on $223{,}834$ GAGLE documents (Alabdulmohsin & Steiner, 2025) scored under Gemma-2B and Mistral-7B. At the population level, all 20 (domain, generator, exponent) comparisons separate human from machine text with non-overlapping 95% intervals. At the document level, a split-half check registered before any detection experiment gives $\rho = 0.063$ for $\widehat{S}$ and 0.576 for $\widehat{H}$ against a bar of 0.60. None of the twenty candidate quantities clears it. Hence μ separates the classes while $\mathrm{Var}(\eta) \gg \mathrm{Var}(\delta)$ at $N \approx 400$ tokens: $\widehat{H}_d$ moves with the population a document came from rather than with the document itself.

Primary Contributions. The fractal signature is a property of the *generating process*, not of a document. (C1) We formulate the estimand/support distinction and pair it with a measurement that separates the two on identical documents (§4.1). (C2) We test four predictions at independent measurement points: three negative predictions and one positive. (C3) *Perturbation as a structural probe*: partitioning an edit battery by index versus realization predicts the *sign* of the family's contribution in all six cases (§4.3). (C4) We derive an admissibility test: split each document's tokens once, re-estimate on each half, and require the correlation to clear a declared bar. The test runs before any detector is trained, and its verdict is independent of the downstream learner.

Secondary Contributions. (S1) *Scorer–attacker lineage inflates measured robustness*, a curvature detector's recall *rises* under paraphrase from its own scorer's lineage and collapses under any other (§4.4). (S2) *Perturbation-response fusion requires editing both classes*. When only machine documents are perturbed, the drift feature, which measures how far a document's features move under an edit, exists only for the positive class. A classifier can then use the feature's presence as the label rather than learning anything about robustness.

The primary finding is not a feature-selection effect. It is *learner-invariant*: a shallow MLP reproduces the pattern and fails more strongly. It is also *scorer-architecture-invariant*, with $\rho = 0.063$ and 0.050 from a 2B Gemma and a 7B Mistral on the same documents.

2 Related Work

Long-range correlation in text is an old observation (Ebeling & Pöschel, 1993; Altmann et al., 2012), formalised by Alabdulmohsin et al. (2024) at $H \approx 0.7$ and extended to model output by Alabdulmohsin & Steiner (2025), whose narrow- versus wide-band contrast is our starting point. Parallel lines measure self-similarity by correlation dimension (Du & Tanaka-Ishii, 2025), spectral scaling (Yang et al., 2026), and the persistent-homology dimension of a token-embedding cloud (Tulchinskii et al., 2023; Wei et al., 2025), the closest neighbours to this work. Their remedy for short text, which is to enlarge the support, is the diagnosis we run in reverse (full treatment, Appendix B). *That short-series scaling estimators are biased is established and is not our contribution* (Cannon et al., 1997; Millán et al., 2021; Grech & Pamuła, 2013; Weron, 2002). Our estimator is the one

this literature deploys, applied to the document lengths practitioners actually score; its instability at short support is therefore the object of the measurement rather than an obstacle to it. What that instability implies for treating the estimate as a per-document *representation* is a reliability question in the psychometric sense (Pronk et al., 2022; Ellis, 2023), and we find no coefficient of that kind reported for a scaling estimator at document length.

3 METHODS

3.1 SURPRISAL TRAJECTORY AND SCALING ESTIMATORS

Let q be a fixed scoring language model and $x = (x_1, \ldots, x_N)$ a document. The surprisal process and its integrated trajectory are

$$s_t \;=\; -\log q(x_t \mid x_{<t})\,, \qquad X_n \;=\; \sum_{t=1}^{n} \frac{s_t - \mu_s}{\sigma_s}, \tag{1}$$

with μ_s, σ_s the document's own mean and standard deviation, so X has empirically centered, standardized increments. For the Hurst exponent, rescaled-range analysis gives $\mathbb{E}\big[R(w)/\sigma(w)\big] \sim w^H$ for window length w, and $\widehat{H}$ is the OLS slope of $\log(R(w)/\sigma(w))$ against $\log w$. Here $H = \frac{1}{2}$ denotes memoryless behavior and $H \to 1$ denotes strong persistence in the limit, including verbatim repetition. For the Hölder exponent, with fixed $\varepsilon \ll 1$ and $p_\varepsilon(\tau) = \Pr\big[|X_{t+\tau} - X_t| \le \varepsilon\big]$, $p_\varepsilon(\tau) \sim \tau^{-S}$, and $\widehat{S}$ is the negative slope of $\log p_\varepsilon(\tau)$ against $\log \tau$. Smaller S means that displacements remain within ε for longer. Both are fit over $\tau \in \{8, 16, 32, 48, 64, 96, 128, 160, 192, 256, 320\}$ with $\varepsilon = 10^{-2}$, calibrated to Eq. 1's z-scoring; per document we drop the first 64 tokens as warm-up, clip to 464, and require 320 surviving tokens, dropping shorter documents rather than padding.

3.2 ESTIMAND, SUPPORT, AND VARIANCE DECOMPOSITION

Both definitions are asymptotic. A document supplies about 400 usable tokens, over which the scale sweep spans more than 1.6 decades. This gap separates two questions that are easily conflated: whether two towns differ in climate, and what one day's reading tells us. Q1 (ensemble): does H, or S, differ between the human and machine generating processes? Pooling realizations answers it. Q2 (support): does $\widehat{H}(x)$ resolve anything about the individual document x beyond the population it came from? That is a question about estimator variance at $N \approx 400$. The two questions are logically independent: Q1 can be yes while Q2 is no. However, a positive Q1 is routinely treated as permission to use the exponent as a document-level feature. That inference is what this paper tests. Writing the per-document estimate for document d with class c_d, domain g_d and temperature T_d:

$$\widehat{H}_d \;=\; \underbrace{\mu(c_d, g_d, T_d)}_{\text{population term}} + \underbrace{\delta_d}_{\text{document term}} + \underbrace{\eta_d}_{\text{estimator error}}, \qquad \mathbb{E}[\eta_d] = 0. \tag{2}$$

In detector-feature terms, μ is the population signal (the generating process that wrote the document), and η is the estimator noise from fitting a slope on ≈ 400 tokens. A reliability coefficient describes the fraction of variance attributed to signal under a measurement-error model. The raw split-half correlation, its Spearman-Brown transformation, and the signal-to-noise ratio are related but distinct quantities. Q1 asks whether μ separates over c; Q2 asks about $\mathrm{Var}(\delta)/\big(\mathrm{Var}(\delta) + \mathrm{Var}(\eta)\big)$. Two halves of the same document also share the population from which it came, so in a corpus whose documents differ in class and domain, the pooled half-half correlation estimates $\mathrm{Var}(\mu{+}\delta)/(\mathrm{Var}(\mu{+}\delta) + \mathrm{Var}(\eta))$, which is strictly the larger quantity (Ellis, 2023). Every pooled coefficient we report is therefore an upper bound on what Eq. 2 asks about a direction generous to the fractal features; we additionally measure the within-population coefficient directly (§3.5).

3.3 HYPOTHESES: FOUR PREDICTIONS OF THE DECOMPOSITION

In the picture of §1, the climate is $\mu(c_d, g_d, T_d)$, the individual house is δ_d, and the thermometer's error is η_d. If $\mathrm{Var}(\eta) \gg \mathrm{Var}(\delta)$, the per-document estimate records which population a document came from rather than what is in it. The rival account is that the fractal family carries genuine

document-level content that is merely *redundant*, swamped by perplexity and style rather than absent. Both predict a weak family; they differ at four points, each measured at an independent point in the pipeline, and each closing an escape route left open by the previous one.

(P1) A second thermometer, told the country already. A detector that already identifies the population by other means gains nothing from the family: deleting all 54 fractal columns from the 85-feature detector should leave its operating point where it was. *(P2) The needle that twitches but says nothing.* Both accounts predict P1, but they differ in how a greedy learner treats the family during training. A genuine-but-redundant signal should receive little modeling attention once cheaper alternatives exist; a population-conditional mean plus noise remains locally correlated with the label and keeps attracting splits it cannot cash out. Split gain should therefore be high while permutation importance remains near zero. *(P3) Two houses on the same street.* With the population term balanced by design across the labels being separated, a genuine document-level signal should appear in full. This is the last place it could hide. *(P4) Move country, and the same thermometer is useful again.* An account that only predicts negative results is not falsifiable; it must also identify a setting where the instrument *does* work. Transferring to a new corpus changes the populations while leaving each document untouched, and the family should regain value because the climate has moved, not because any document is read more accurately.

3.4 Data and Features

All measurements use GAGLE (Alabdulmohsin & Steiner, 2025): five domains crossed with two generators (Gemma-2B, Mistral-7B), three temperatures, and eight prompt types, together with the human source articles on which those generations were conditioned. Documents shorter than the token window are excluded, leaving $223{,}834$: $112{,}402$ from Gemma-2B, $106{,}432$ from Mistral-7B, and $5{,}000$ human, a human prevalence of 2.234%. Each document carries 85 features in three families (Appendix D). Family A, the 54 fractal features, holds both exponents together with their fit quality, uncertainty, per-scale residuals, and windowed profiles. Family B, 17 token-statistical features, holds log-perplexity moments, rank buckets (Gehrmann et al., 2019), and entropy. Family C, 14 discourse and style features, holds length, repetition, and dispersion statistics computed on a ratio-derived word budget, which keeps them from degenerating into length proxies. GAGLE's precomputed scores record neither the checkpoint nor the tokenizer that produced them, so Family B is populated only where a document's own rescoring aligns with the released values at $r_{\text{align}} \geq 0.98$. For P4, RAID (Dugan et al., 2024) (11 generators, 8 domains) is rescored through the identical pipeline and used only as an external test set.

Corpus composition. Two properties of GAGLE constrain how the detection results should be read. First, the classes are strongly imbalanced. Human text forms 2.234% of the retained corpus, so human-side intervals are wider than machine-side intervals, the learner is class-weighted, and detection performance is reported as recall at fixed false-positive rates rather than as accuracy. Precision in deployment depends on the prevalence of machine-generated text, which is usually far lower than in GAGLE (Table 12). Second, document length differs by class by construction. After the token window is applied, every human document retains exactly 400 usable tokens, whereas machine documents retain between 320 and 464. Usable length alone therefore separates the classes in part, and every detection result is read against a two-feature length control (§4.2.1). The reliability analysis is not exposed to this confound in the same way, because both halves of a split come from the same document and share its length. Within each usable-length band, none of the four headline quantities reaches the 0.60 bar.

3.5 Split-Half Reliability Protocol

The check and its threshold were fixed before any detection experiment was run. Each document's token sequence is split into odd and even halves, and both exponents are refit independently on each half. The correlation between the two fits is then Spearman–Brown corrected, $\rho = 2r/(1 + r)$, against a bar of $\rho \geq 0.60$ declared in advance. Interleaving is the primary construction because a contiguous cut confounds how much the estimator varies on repeated measurement of the same text with how much a document genuinely drifts in style between its opening and its close; seven further constructions, from blocks of increasing size to a contiguous cut, are run as sensitivity checks (why decimation preserves the estimand, Appendix F.2). The within-population coefficient ρ_{within}

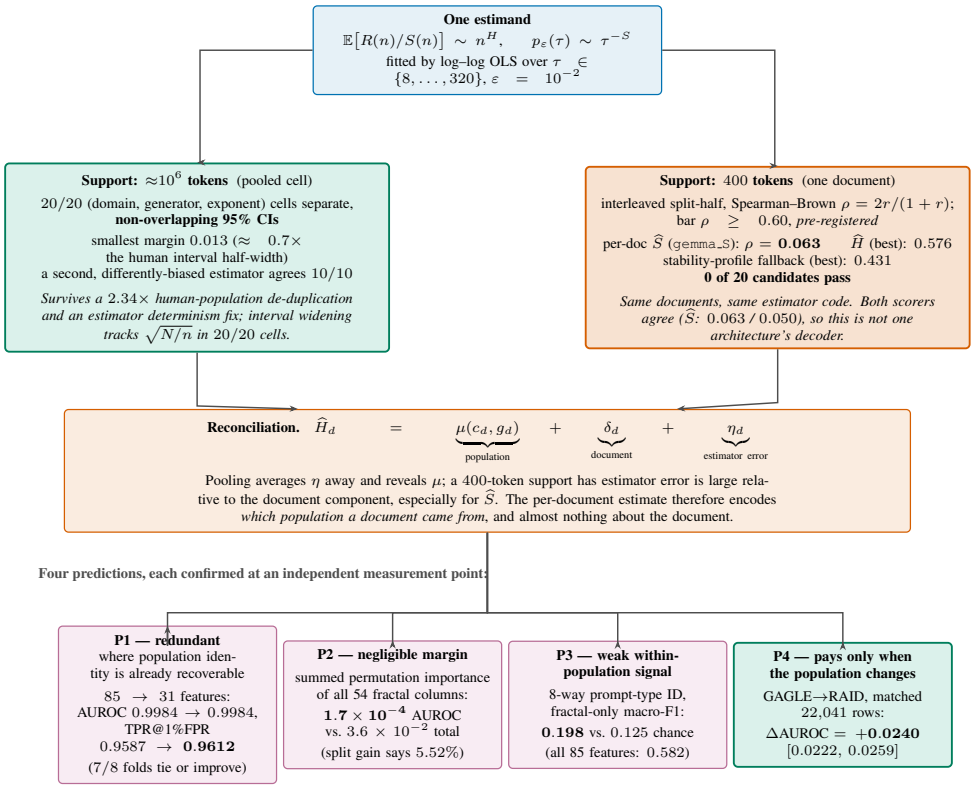

Figure 1: One estimand, two supports. *Left:* pooled, the exponents separate in $20/20$ cells with non-overlapping intervals. *Centre:* the same documents at 400-token support give $\rho = 0.063$ for $\widehat{S}$ and $0/20$ features clearing $\rho \geq 0.60$. *Right:* the four predictions Eq. 2 makes each confirmed independently.

demeans each half by its own (c, g) stratum before correlating, which removes μ by construction. The support needed to clear the bar, $N^\star$, is obtained from an estimator noise floor calibrated on synthetic fractional Gaussian noise (Appendix F.4).

3.6 Detection Instrument and Statistical Analysis

The learner is deliberately unremarkable and held fixed across comparisons: a class-weighted gradient-boosted tree ensemble with splits frozen before training, and groups disjoint by the seed article. Eight folds cross five confound configurations, arms, and learners (304 cells, 2,000-fold bootstrap intervals); the headline metric is recall at fixed low FPR, not AUROC. Feature families are priced by ablation arms, by split gain, and by held-out permutation importance. P3 is an eight-way prompt-type identification on machine-generated rows only, scored by macro-F1. The structural probe applies six edit operators to the GAGLE machine text, partitioned by what they destroy (§4.3), each pair passing a semantic-preservation guard (Appendix J).

4 Results

4.1 Answers to Q1 and Q2

4.1.1 Ensemble Gap Holds (Q1)

Two artifacts could produce a corpus-level separation even if the generating processes did not differ, and both were corrected before any interval below was computed. The first is duplication: GAGLE's split label forms part of an article's identity, so collapsing recurring articles takes the human N from $\approx 11{,}700$ to $5{,}000$ and widens the human intervals accordingly. The second is esti-

mator non-determinism: the Hölder estimator sampled window starts rather than enumerating them; enumeration removes the resulting Monte-Carlo noise.

The separation persists in all 20 cells. The narrowest margin anywhere is `wikipedia` S under Gemma-2B, a gap of 0.013, about 0.7 of the human interval's half-width (Table 1, Appendix E). Replacing the pooled estimate with the mean of the individual document estimates, and resampling documents to obtain their intervals, separates all ten S cells in the same direction; the two routes carry different biases, so their agreement is not automatic (further checks, Appendix E).

4.1.2 Reliability at Document Support (Q2)

None of the twenty candidate quantities reaches $\rho \geq 0.60$ (full table, Appendix F). $\widehat{H}$ is modestly but genuinely reliable and lands just under the bar: $\rho = 0.576$ for `gemma_H`, 0.547 for `mistral_H`. $\widehat{S}$ has almost none: 0.063 and 0.050. At $N \approx 400$ tokens, one document's Hölder exponent measured on its odd tokens and again on its even tokens agrees about as well as estimates taken from two documents picked at random. $\widehat{H}$'s shortfall exceeds sampling error: at $n = 221{,}904$ its Fisher-z 95% interval is $[0.573, 0.580]$. $\widehat{S}$ misses by an order of magnitude, and the pre-registered fallback moves the wrong way: windowed stability tops out at 0.431 and bottoms out at 0.003. The result holds across a 2B Gemma and a 7B Mistral scorer, and it was measured after the estimator's non-determinism of §4.1.1 was corrected and the corpus regenerated.

Across the eight constructions (a ninth was measured and excluded as invalid, Appendix F.2), reliability declines smoothly and monotonically as the block grows: for `gemma_H`, $0.576 \to 0.562 \to 0.549 \to 0.534 \to 0.501$ at block sizes $1, 4, 8, 16, 32$. If decimation were damaging the estimand, larger blocks would recover reliability; instead, it falls, and single-token interleaving is the most favorable of the eight. Only the contiguous split sits outside the band; the other seven share the signature of stylistic drift that the pre-registration anticipated.

Removing the population term. ρ_{within} barely moves from the pooled value: 0.552 against 0.576 for `gemma_H`, 0.520 against 0.547 for `mistral_H`, and 0.060/0.046 against 0.063/0.050 for $\widehat{S}$. Population structure accounts for under 4% of single-half variance for $\widehat{H}$ and under 0.2% for $\widehat{S}$, so $\widehat{S}$'s near-zero reliability is document-level estimation noise (Appendix F.3).

Refitting the same halves with detrended fluctuation analysis and local Whittle gives a mixed answer: local Whittle clears the bar for both scorers ($\rho = 0.677$ and 0.642), DFA clears it for Gemma only, at the margin. R/S and DFA agree at 0.87 and 0.86, consistent with one exponent measured less noisily; R/S and Whittle agree at only 0.76 and 0.75, so part of Whittle's headroom comes from measuring a different quantity. Neither raises $\widehat{S}$ (Appendix F.5). Against the synthetic fractional Brownian motion of known exponent, $\widehat{S}$ at document support is biased by -0.26 to -0.48; additional bootstrap draws at the same support do not close the gap, which locates the bias in the support rather than in the resampling. Every machine cell sits above its matched human cell in the pooling mode used, so the bias runs against the ensemble gap: the separation of §4.1.1 is understated, not manufactured (Appendix F.1).

$\widehat{H}$ crosses the bar at $N^\star = 800$, roughly twice GAGLE's support, and already reaches $\rho_{\text{SB}} = 0.49$ and 0.46 on the synthetic floor at $N = 400$. $\widehat{S}$ requires $N^\star \geq 6{,}400$, sixteen times what any document here supplies (Figure 3, Appendix F.4); the median full-document $\widehat{H}$ of $5{,}000$ human documents, 0.629, lies inside the calibration band. For $\widehat{S}$ the floor predicts 0.229 and 0.197 at $N = 400$ against 0.063 and 0.050 measured, which makes the quoted $N^\star$ a lower bound. The gate emitted an empty whitelist; the bar stays at 0.60, and the detection stage trains on all 54 fractal features from a manifest declared in advance so that the ablations below price them independently of the gate's verdict.

4.2 Tests of the Four Predictions

4.2.1 P1 and P2, Necessary Conditions

Two controls bound the instrument first (Appendix G): a label-shuffle harness lands systematically *below* chance, and a two-feature `length_only` arm prices the length confound at AUROC 0.8635

and TPR@1%FPR 0.0477. *P1 holds.* Deleting all 54 fractal columns leaves AUROC unchanged to four decimals and moves recall at 1% FPR from 0.9587 to 0.9612, a 31-feature model matching an 85-feature one, with the movement in the *wrong direction* for a family supposed to be carrying a signal. A fractal-only model reaches AUROC 0.6629, so the family is not inert; `fractal+stat` sits *below* `stat_only`, so its contribution is not additive; `all_minus_fractal` ties or wins on 7 of 8 frozen folds, and a shallow MLP reproduces the pattern.

P2 holds. Split gain over all 85 inputs credits the 54 fractal columns with 5.52% of the total, and the model splits on every one of them at least 125 times. On the held-out fold, the summed AUROC drops across all 54 is $\mathbf{1.7 \times 10^{-4}}$, against 1.54×10^{-2} and 2.04×10^{-2} for the other two families: 0.47% of the total against a 5.52% gain share, with 35 of 54 at zero or negative (Appendix H). The family is used constantly, but contributes almost nothing out of sample.

4.2.2 Critical Test (P3)

GAGLE is a complete 240-cell factorial (2 generators $\times$ 3 temperatures $\times$ 5 domains $\times$ 8 prompt types), so in eight-way prompt-type identification on machine rows, μ varies across documents but is balanced across the labels being separated by the corpus's own design. Chance is macro-F1 0.125. All 85 features reach 0.5816 [0.5747, 0.5885], so the eight classes genuinely differ document by document; the 54 fractal features alone reach 0.1981 [0.1919, 0.2043], about a sixth of the distance from guessing to what the corpus supports, despite having 1.7$\times$ as many columns as the incumbents. Three checks close the residual-confound route (Table 16, Appendix I.1): a population-only control (generator, domain, temperature, no prompt type) scores *below* chance (0.0594); the factorial survives the token-window and complete-case gates with a maximum cell deviation of 0.22 percentage points; and removing each fractal column's stratum mean before refitting leaves fractal-only macro-F1 unchanged, its interval overlapping ($0.1981 \rightarrow 0.2046$). The residual that survives these checks is therefore attributable to the document term rather than to population leakage, but it is small: the fractal family covers about one-sixth of the distance between chance and the full-feature score on the same task. P3 thus establishes that δ is non-zero, not that it is absent; its magnitude relative to η is fixed by the reliability measurements of §4.1.2. Two further within-corpus results separate the same two terms (Appendix G): on *substitution*, a question about δ, the fractal arm loses on all eight folds against `stat_only`; *generator identity*, a question about μ, reaches AUROC 0.9943 with the same features. The detector's actual failure is not a fractal-specific problem; it is a tail-ranking failure on `billsum` LODO driven by prompt type, whose largest shifts are style features.

4.2.3 Where Ensemble Structure Does Transfer (P4)

Applied to RAID with *no* retraining, crossing corpora costs 0.195 AUROC ($0.9984 \rightarrow 0.8032$) and 0.731 recall at 1% FPR ($0.9587 \rightarrow 0.2277$). The largest degradation is anywhere in this evaluation. On a composition-matched population, the fractal family now contributes a small but statistically significant +0.0240 AUROC (Table 14, Appendix I). Tested directly, replacing each document's 54 fractal columns with its (generator, domain) group mean *recovers more than the entire* gap (AUROC 0.8149, 149% of it), and subtracting it retains 59% (AUROC 0.7934) (Appendix I.2). `fractal_only` does not transfer.

4.3 Structural Probe

The scaling estimators of §3 are functionals of the ordered token index, indifferent to *which* words appear. The battery splits edits along that axis: Class I (order), `sentence_reorder` and `typo_noise`, scrambles *where* tokens sit, not *which* words they are; Class II (content) contains four paraphrase, style, and summarization operators that swap *which* words appear while preserving coarse-scale order (Appendix J). A document-level property should be most affected by Class I; population identity plus estimator noise should instead make the family actively *harmful* under Class I.

The six operators yield 39,810 (document, variant) pairs. The tabular detector (Table 18) stays above 0.92 on every Class II operator and collapses to 0.2670 on `sentence_reorder`, while Binoculars (Hans et al., 2024) degrades roughly uniformly, and a fine-tuned RoBERTa is unmoved. Retraining the fractal-free 31-feature model on the same edited population, the partition predicts the sign in all

six cases: the full model is ahead by 0.94–2.10 points on all four Class II operators, and *deleting* the fractal family *raises* recall by 1.96 and 8.76 points on `sentence_reorder`/`typo_noise`.

4.4 Scorer-Attacker Lineage Inflates Measured Robustness

This result uses the same edit battery but does not concern fractal features, and nothing in §4.1–§4.3 depends on it. Fast-DetectGPT's recall *rises* under every Class II operator, from a 0.5104 baseline to 0.63–0.86, paraphrasing making machine text easier to detect, the opposite of this literature's central finding (Krishna et al., 2023; Sadasivan et al., 2025). It is neither a floor artifact (far above our 15% no-signal screen) nor robustness: Fast-DetectGPT scores curvature under Gemma-2B, and these paraphrases were generated in Gemma's own lineage, so the edited text is maximally predictable *to the scorer*. Regenerating the same paraphrase outside that lineage collapses recall to 1.2–4.5% across a capability ladder of five external models on $5{,}000$ documents. At the same time, the tabular detector degrades broadly monotonically with attacker size ($72.6\% \to 36.5\%$) and Binoculars traces the cleanest curve ($70.8\% \to 35.1\%$; Appendix K). Scorer–attacker lineage, not detector design, determines whether a curvature-based detector appears robust to paraphrase.

5 Discussion

The four predictions of §3.3 hold, and they are predicted by the reliability measurement rather than merely consistent with it. Without $\rho = 0.063$, a family that separates human from machine text in 20 of 20 cells yet contributes nothing to a classifier on the same corpus looks like a modeling failure, and the natural response is an architecture search. It is not one: the estimate is a population-conditional mean plus noise, perplexity, and style already identify the population, so the fractal columns arrive carrying information the model has, wrapped in variance it does not want (P1). Noise that correlates with the label at the population level keeps looking locally splittable without generalizing, which is the gain–importance divergence of P2. P3 removes the cover that P1 and P2 leave open: with μ balanced by design, the same features that recover generator identity at AUROC 0.9943 recover a sixth of the achievable prompt-type signal. The claim is accordingly one of magnitude: the document term exists, but is dominated by estimator error at this support, so $\mathrm{Var}(\eta) \gg \mathrm{Var}(\delta)$ should not be read as $\delta = 0$. P3 is a within-population discriminative test, not a direct estimate of $\mathrm{Var}(\delta)$, and it does not reference the split-half gate, so $\rho = 0.063$ should be read as a property of the estimand–support relationship rather than of how tokens were split.

`sentence_reorder` and `typo_noise` are the only operators that attack the *support* rather than the *content*, preserving the lexical statistics every other feature reads and perturbing only the index Family A claims to measure. A representation that genuinely encoded that structure would degrade under them; instead, its contribution *flips sign* precisely and only there. A feature family cannot measure sequence structure and also be actively harmful when sequence structure is perturbed. This is what $\mathrm{Var}(\eta) \gg \mathrm{Var}(\delta)$ predicts, and it isolates δ without relying on the split-half gate.

The $+0.0240$ on RAID would contradict the central claim only if it were document-level. What moved is μ: the group-mean replacement recovers 149% of the gain, the gain appears only after the model has lost 0.195 AUROC crossing corpora, and `fractal_only` does not transfer. Population information does the majority of the work, although the 59% retained after subtraction indicates a real document-level residual. The ensemble term is the part that travels.

Nothing here contradicts the language's fractal information structure or model output's departure from it. An exponent defined by an asymptotic law and estimated on a 400-token realisation across 1.6 decades resolves the process far better than the realization, and using the whole curve does not rescue the estimate. Wei et al. (2025) find a topological scaling dimension on *embeddings* degrading at short length for the same reason, and Yang et al. (2026) report a $5/3$ spectrum disrupted by randomization of token order. On a representation with no fractal machinery, overlapping-window MiniLM embeddings reduced to persistent homology loop statistics at zero content overlap between halves, no window geometry tried clears $\widehat{H}$'s own reliability , evidence for the generality of the mechanism beyond this estimator family. The lineage result of §4.4 is a second instance of the same pattern: a measurement appears to read the document in front of it while reading an artifact of the evaluation design, here the attacker's kinship with the scorer. The check is inexpensive: regenerate one attack type outside the scorer's lineage and compare.

The interpretation is regime-dependent. Within a known corpus, the family is redundant at best and costs 8.76 points of recall at worst; for a detector that must generalize, it contributes $+0.0240$ AUROC. The practical recommendation is to keep these columns when the population is expected to change, but not to treat them as descriptions of the document in front of you. *No detector evaluated here supports consequential accusations against individuals*: the best detector loses 73 points of recall when crossing corpora.

5.1 Limitations

Several limitations bound these conclusions. First, the edit battery perturbs only machine-generated text ($9{,}997$ of $9{,}997$ edited documents), so a perturbation-response (drift) feature exists for the positive class alone, and a classifier trained on it would learn class membership rather than robustness. Perturbation-response fusion therefore requires a class-symmetric battery that edits human and machine text alike; we state this as a protocol requirement rather than evaluate it. Second, the evidence rests on a single scorer pair and one English corpus family in which the human class is both rare (2.23%) and fixed in length. Length-matched bands (§3.4), a complete-case RAID population covering 32% of its rows (§4.2.3), and prevalence-adjusted precision (Table 12) mitigate these constraints without removing them, and a second representation family, persistent homology of embedding trajectories, was examined only as a pilot (§5). Third, we offer no architectural account of the exponents and do not cover human revision of machine text. Fourth, the 65 comparisons of the detection grid are reported without multiplicity correction. Finally, two feature-extraction defects were found during review; every feature store was rebuilt, and every number reported here is post-correction, with one marked exception (Table 18).

6 Conclusion

The fractal signature of machine-generated text is real, but at this support, it is not a document representation. All 20 ensemble cells separate, none of the 20 split-half candidates clears the pre-registered bar at 400 tokens, and all four predictions of §3.3 hold. We therefore propose a measurement-validity criterion, not a theorem: a scaling statistic should not be treated as a stable document-level measurement without a reliability coefficient evaluated at the support on which it is used.

Use of AI Assistance

We disclose the use of large language models (LLMs), principally Claude (Anthropic), in the preparation of this work. LLMs were used to edit and refine prose and to draft preliminary text, which the authors subsequently rewrote and verified; to search for, retrieve and summarize related work, with every reference opened and confirmed by the authors before citation; to support research ideation and execution, including discussion of candidate analyses and controls and the writing and debugging of analysis code; and to check mathematical claims, including the variance decomposition and the reliability derivations, all of which the authors re-derived independently. Final decisions on hypotheses, experimental design and interpretation were made by the authors, and every reported number was checked against the artifact that produced it. Open-weight models used as experimental objects (scorers and attack models) are part of the methodology and are described in the main text. The authors take full responsibility for the content of this paper.

Ethics Statement

This work studies detectors of machine-generated text, a dual-use subject: the operator taxonomy of Section 4.3 informs defenders and attackers alike. We consider publication net positive, because the vulnerabilities concern an already-published feature family and trivially available edit operators, the practical recommendation is to remove an unreliable component, and the scorer–lineage finding (Section 4.4) corrects an evaluation practice that overstates robustness. Because false accusation is the principal harm, headline results are reported as recall at 1% and 5% false-positive rate, and no evaluated detector is presented as suitable for consequential decisions about individuals (Section 5.1); false-positive rates for specific writer populations, including non-native writers, were not

measured. Only public research corpora (GAGLE, RAID) were used; no human-subjects data or annotation was collected.

Reproducibility Statement

Code, frozen artefacts, and step-by-step instructions are available in the attached supplementary. A CPU-only check, which runs in about two minutes without downloading data, regenerates the synthetic noise-floor experiment and the reliability tables and compares them numerically with the released reference outputs.

References

Ibrahim Alabdulmohsin and Andreas Peter Steiner. A tale of two structures: Do LLMs capture the fractal complexity of language? In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 964–1025. PMLR, 2025. arXiv:2502.14924.

Ibrahim Alabdulmohsin, Vinh Q. Tran, and Mostafa Dehghani. Fractal patterns may illuminate the success of next-token prediction. *Advances in Neural Information Processing Systems*, 37: 112864–112888, 2024. arXiv:2402.01825.

Eduardo G. Altmann, Giampaolo Cristadoro, and Mirko Degli Esposti. On the origin of long-range correlations in texts. *Proceedings of the National Academy of Sciences*, 109(29):11582–11587, 2012.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations (ICLR)*, 2024.

Mary J. Cannon, Donald B. Percival, David C. Caccia, Gary M. Raymond, and James B. Bassingthwaighte. Evaluating scaled windowed variance methods for estimating the Hurst coefficient of time series. *Physica A: Statistical Mechanics and its Applications*, 241(3):606–626, 1997.

Yize Cheng, Vinu Sankar Sadasivan, Mehrdad Saberi, Shoumik Saha, and Soheil Feizi. Adversarial paraphrasing: A universal attack for humanizing AI-generated text. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.07001.

Xin Du and Kumiko Tanaka-Ishii. Correlation dimension of auto-regressive large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2510.21258.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. arXiv:2405.07940.

Werner Ebeling and Thorsten Pöschel. Entropy and long-range correlations in literary English. *Europhysics Letters*, 26(4):241–246, 1993.

Jules L. Ellis. The Spearman–Brown formula and reliabilities of random test forms. *arXiv preprint arXiv:2308.13811*, 2023. arXiv:2308.13811.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. GLTR: Statistical detection and visualization of generated text. In *Proceedings of ACL 2019, System Demonstrations*, pp. 111–116, 2019.

Dariusz Grech and Grzegorz Pamuła. On the multifractal effects generated by monofractal signals. *Physica A: Statistical Mechanics and its Applications*, 2013. arXiv:1307.2014.

Yikai Guo, Bin Wang, Xilai Fan, Wenjun Ke, and Haoran Luo. On the salience of low-probability tokens for AI-generated text detection: A multiscale uncertainty perspective. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2606.02158.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2401.12070.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2303.13408.

Lucio La Cava and Andrea Tagarelli. Luminol-AIDetect: Fast zero-shot machine-generated text detection based on perplexity under text shuffling. *arXiv preprint arXiv:2604.25860*, 2026.

Zhichen Liu, Kaitong Qin, Linhan He, and Yang Xu. WaveDetect: Robust framework for machine-generated text detection via wavelet transform. *arXiv preprint arXiv:2606.23336*, 2026.

Ning Lu, Shengcai Liu, Rui He, Qi Wang, Yew-Soon Ong, and Ke Tang. Large language models can be guided to evade AI-generated text detection. *Transactions on Machine Learning Research*, 2024. arXiv:2305.10847.

Sina Mansouri, Mohit Marvania, and Abolfazl Safikhani. Stability-aware feature design for robust watermark detection in machine-generated text. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2608.18102.

Clara Meister and Ryan Cotterell. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 5328–5339, 2021.

Ginno Millán, Román Osorio-Comparán, and Gastón Lefranc. Preliminaries on the accurate estimation of the Hurst exponent using time series. *arXiv preprint arXiv:2103.02091*, 2021. arXiv:2103.02091.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning (ICML)*, 2023.

Thomas Pronk, Dylan Molenaar, Reinout W. Wiers, and Jaap M. J. Murre. Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. *Psychonomic Bulletin & Review*, 29:44–54, 2022. doi:10.3758/s13423-021-01948-3.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-generated text be reliably detected? Stress testing AI text detectors under various attacks. *Transactions on Machine Learning Research*, 2025. arXiv:2303.11156.

Mihir Sahasrabudhe. Reversal invariance in autoregressive language models. *arXiv preprint arXiv:2511.00341*, 2025.

Shuntaro Takahashi and Kumiko Tanaka-Ishii. Evaluating computational language models with scaling properties of natural language. *Computational Linguistics*, 45(3):481–513, 2019.

Lekkala Sai Teja, Annepaka Yadagiri, Sangam Sai Anish, Siva Gopala Krishna Nuthakki, and Partha Pakray. Modeling the attack: Detecting AI-generated text by quantifying adversarial perturbations. *arXiv preprint arXiv:2510.02319*, 2025.

Brian Tufts, Xuandong Zhao, and Lei Li. A practical examination of AI-generated text detectors for large language models. *arXiv preprint arXiv:2412.05139*, 2024.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Serguei Barannikov, Irina Piontkovskaya, Sergey Nikolenko, and Evgeny Burnaev. Intrinsic dimension estimation for robust detection of AI-generated texts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.04723.

Dongjun Wei, Minjia Mao, Xiao Fang, and Michael Chau. Short-PHD: Detecting short LLM-generated text with topological data analysis after off-topic content insertion. *arXiv preprint arXiv:2504.02873*, 2025.

Rafał Weron. Estimating long-range dependence: finite sample properties and confidence intervals. *Physica A: Statistical Mechanics and its Applications*, 312(1–2):285–299, 2002. arXiv:cond-mat/0103510.

Alva West, Yixuan Weng, Minjun Zhu, Luodan Zhang, Zhen Lin, Guangsheng Bao, and Yue Zhang. AI-generated text is non-stationary: Detection via temporal tomography. *arXiv preprint arXiv:2508.01754*, 2025.

Zhongxin Yang, Chun Bao, Yuanwei Bin, Xiang I. A. Yang, and Shiyi Chen. Turbulence-like 5/3 spectral scaling in contextual representations of language as a complex system. *arXiv preprint arXiv:2604.05536*, 2026.

A NOTATION

Every mathematical symbol used in this paper. Hatted symbols are always *estimates* computed on a finite document; bare symbols are the corresponding *estimands*, properties of the generating process defined by an asymptotic limit.

Corpus, scoring, and the surprisal trajectory.

Symbol	Meaning
q	The fixed scoring language model (Gemma-2B or Mistral-7B)
$x = (x_1, \ldots, x_N)$	A document, as a token sequence
x_t	The token at position t
$x_{<t}$	All tokens preceding position t
s_t	Surprisal of token t, $-\log q(x_t \mid x_{<t})$
$\mu_s,\ \sigma_s$	Mean and standard deviation of *that document's own* surprisal series
X_n	Integrated trajectory, cumulative sum of empirically centered, standardized surprisal increments
N	Usable token support of one document (≈ 400 after the warm-up, clip and floor gates)
d	Document index
$c_d,\ g_d,\ T_d$	Class (human/machine), domain, and decoding temperature of document d
k	A (generator, domain) group

The two scaling exponents, and the decomposition of a per-document estimate.

Symbol	Meaning
H	Hurst exponent the *estimand*, defined by $\mathbb{E}[R(w)/\sigma(w)] \sim w^H$ as $w \to \infty$. $H = \frac{1}{2}$ is memoryless; $H \to 1$ is strong persistence
$\widehat{H}$	The per-document Hurst *estimate*, OLS slope of $\log(R(w)/\sigma(w))$ against $\log w$
S	Hölder exponent the estimand, defined by $p_\varepsilon(\tau) \sim \tau^{-S}$
$\widehat{S}$	The per-document Hölder estimate; the negative slope of $\log p_\varepsilon(\tau)$ against $\log \tau$
w	Window length in rescaled-range analysis
$R(w)$	Range of the cumulative deviation within a window of length w
$\sigma(w)$	Standard deviation within that window (the "S" of R/S, renamed to avoid collision with S)
τ	Lag, fit over $\tau \in \{8, 16, \ldots, 320\}$
ε	Displacement threshold for the Hölder fit, $\varepsilon = 10^{-2}$, calibrated to Eq. 1's z-scoring
$p_\varepsilon(\tau)$	$\Pr[\lvert X_{t+\tau} - X_t\rvert \le \varepsilon]$
$\mu(c_d, g_d, T_d)$	Population term, the class-, domain- and temperature-conditional mean. The "climate"
δ_d	Document term, what is specific to document d. The quantity a document-level representation would have to carry
η_d	Estimator error at finite support, $\mathbb{E}[\eta_d] = 0$. The thermometer's own noise
$\mathrm{Var}(\cdot)$	Variance; the object of interest is $\mathrm{Var}(\delta)/(\mathrm{Var}(\delta) + \mathrm{Var}(\eta))$

Reliability, support, and evaluation.

Symbol	Meaning
r	Raw correlation between a document's two half-estimates
ρ_{SB}	Spearman–Brown corrected reliability, $\rho_{\text{SB}} = 2r/(1+r)$; equivalently, the fraction of variance that is signal rather than noise
ρ_{pooled}	Reliability pooled over a corpus whose documents differ in class and domain; an *upper bound*, since μ enters the numerator
ρ_{within}	Reliability after demeaning each half by its own (c, g) stratum, removing the population term by construction
$N^\star$	Support length at which ρ_{SB} crosses the pre-registered 0.60 bar
b	Block size in the split construction; $b = 1$ is single-token interleaving
n	Number of documents entering a given correlation
$t_1,\ t_5$	Decision thresholds at 1% and 5% false-positive rate
r_{align}	Alignment gate above which Family-B features are written (≥ 0.98)
$\bar{x}_k$	Mean of a fractal column within group k, used to price the population term in transfer
$N_{\text{pre}},\ N_{\text{post}}$	Human document counts before and after de-duplication; theory predicts intervals widen by $\sqrt{N_{\text{pre}}/N_{\text{post}}}$
Φ	Standard normal CDF, used when restating Wei et al. (2025)'s detection probability

B Extended Related Work

Long-range correlation in text is an old observation (Ebeling & Pöschel, 1993; Altmann et al., 2012), formalised for information-theoretic complexity by Alabdulmohsin et al. (2024) with $H \approx 0.7$ and extended to model output by Alabdulmohsin & Steiner (2025), whose narrow-band/wide-band contrast is our starting point. Parallel lines measure self-similarity by correlation dimension over next-token log-probability vectors (Du & Tanaka-Ishii, 2025), by scale-free property tests (Takahashi &

Tanaka-Ishii, 2019; Meister & Cotterell, 2021), by spectral scaling of contextual embedding trajectories (Yang et al., 2026), and by the persistent-homology dimension of a document's token-embedding cloud (Tulchinskii et al., 2023; Wei et al., 2025). *We differ* in the question, not whether the scaling structure is present or informative in the ensemble, we confirm it is, but whether the estimator recovering it has enough support on one document to say anything about that document. To our knowledge, no prior work reports a reliability coefficient for these estimators at the document length.

On the detection side, log-probability thresholds and rank visualization (Gehrmann et al., 2019), curvature-based zero-shot scores (Mitchell et al., 2023; Bao et al., 2024), cross-model perplexity ratios (Hans et al., 2024) and fine-tuned encoders are the incumbents against which any additional feature family must be priced. West et al. (2025) argue that aggregating token-level measurements into a scalar discards positional information; our result is a complementary caution, retaining the whole scaling curve does not help either, because the curve is estimated on the same short support. *We differ* in treating the fractal family as an *additive* component priced against these incumbents rather than as a replacement, and in pricing it by permutation importance on held-out folds rather than by split gain.

Paraphrase defeats most detectors (Krishna et al., 2023), recursive paraphrase more so (Sadasivan et al., 2025), and prompt- or policy-optimised attackers more so again (Lu et al., 2024; Cheng et al., 2025); RAID (Dugan et al., 2024) is the standard shared benchmark for evaluating these claims across generators, domains, and attacks. *We differ* in using the edit battery as a *structural probe* rather than as a leaderboard, the operators are partitioned a priori by what they act on, and the partition is used to predict the sign of a feature family's contribution, which is a different use of an attack cascade rather than measuring how much recall it removes.

Several recent evaluations find detector performance far below headline claims once the evaluation population is not the training population (Tufts et al., 2024; Dugan et al., 2024). *We differ* in supplying a mechanism, the reliability coefficient of §4.1 bears on *why* a particular representation does not transfer, and was taken before the transfer experiments were run, rather than as an observation that it did not.

Finally, the operator partition of §4.3 is a general construction rather than a property of this feature family, for any representation defined as a functional of some structure in the input, an operator acting on that structure, and one acting around it form a two-sided test. We would expect it to separate positional from non-positional detectors (West et al., 2025; La Cava & Tagarelli, 2026) and applied to a scoring model rather than a detector, to bear on the reversal-invariance property of the causal language-modelling objective (Sahasrabudhe, 2025).

C What Prior Work Says About Its Own Estimator

The two passages from Tulchinskii et al. (2023) referenced in §1 are quoted in full because this paper's framing depends on them. Their construction is a per-document scaling estimate: they compute the persistent-homology dimension "of each text sample, considering it as a separate manifold" by fitting a log–log linear regression over a sweep of subsample sizes, report a distribution-level separation (9.49 ± 1.01 for human text against ≈ 8 for the generated), and then threshold the per-document estimate as a detector. Their first stated limitation is that the estimator "is stochastic in nature, PH dimensions of texts from the same generator vary widely, and the estimation algorithm is stochastic as well, which adds noise," the document term (δ) and the estimator term (η) of Eq. 2 named in a single sentence, but separated in neither that work nor, to our knowledge, subsequent work. Observing that their misclassified outliers are disproportionately short, they further write that "PH dimension estimation is overall less correct on short texts" and "leave a more comprehensive analysis of such failure cases for future work." The split-half coefficient reported here is one way to run that analysis. It is estimator-agnostic and requires only that the quantity be recomputable on a subsequence.

Wei et al. (2025) confronts the same problem at short length, traces it to a "local density peak" in the same log–log fit, and write the detection probability as $\Phi\big((\mu^h - \mu^m)/\sqrt{(\sigma^h)^2 + (\sigma^m)^2}\big)$. Their σ is the between-document standard deviation, which contains $\mathrm{Var}(\delta) + \mathrm{Var}(\eta)$ and cannot separate them; their remedy is to enlarge the support by inserting off-topic content before each document,

which lowers σ from 1.49 to 0.79 on one subset and 1.54 to 1.00 on another while preserving the class gap.

D FEATURE INVENTORY

The 85 columns are declared in `feature_store/p3_feature_manifest.json`. ***Family A (fractal,*** 54***):*** $\widehat{S}$, $\widehat{H}$, log–log fit R^2 and bootstrap CI half-widths per scorer; per-scale log–log residuals at $\tau \in \{8, 16, 32, 48, 64\}$; intra-document windowed mean, variance, min, max, and drift of both exponents, and the two cross-scorer divergences $|\widehat{S}_{\text{gem}} - \widehat{S}_{\text{mis}}|$ and $|\widehat{H}_{\text{gem}} - \widehat{H}_{\text{mis}}|$. ***Family B (token-statistical,*** 17***):*** log-perplexity mean, variance, skewness and kurtosis per scorer; GLTR-style rank-bucket fractions (Gehrmann et al., 2019); exact next-token entropy; and a cross-scorer perplexity ratio. ***Family C (discourse/style,*** 14***):*** sentence and length statistics, type–token ratio, bigram repetition, sentence-transition entropy, four discourse-marker rates, and MiniLM embedding dispersion, all computed on a per-document ratio-derived word budget, so they cannot degenerate into length proxies.

Four of eight prompt types largely fail the Family-B alignment gate ($r \geq 0.98$) by design; the failure is recorded with an explicit `Alignment_Passed=False` flag and never imputed. Eleven provenance flags and six window-length columns are excluded from the feature matrix entirely, and a `provenance_only` arm prices that exclusion at AUROC 0.5655 (Appendix G).

E CORPUS-LEVEL CHECKS

Table 1: Corpus-level pooled exponents after both corrections; 95% intervals, 10-fold outer bootstrap re-pooling raw trajectories. Human $n = 1{,}000$ per domain, machine $n \approx 21{,}000$–$22{,}700$ per cell, all scored under Gemma-2B. No interval overlaps its domain's human interval on either exponent for either generator, $20/20$.

	Hölder S			Hurst H		
Domain	Human	`gemma_2b`	`mistral_7b`	Human	`gemma_2b`	`mistral_7b`
`big_patent`	[.4135,.4283]	[.4762,.4806]	[.4929,.5012]	[.6365,.6415]	[.6891,.6911]	[.7123,.7141]
`billsum`	[.3992,.4268]	[.4541,.4638]	[.4478,.4538]	[.6315,.6368]	[.6674,.6693]	[.6859,.6879]
`newsroom`	[.3539,.3794]	[.4627,.4668]	[.4888,.4917]	[.6021,.6122]	[.6613,.6635]	[.6893,.6918]
`scientific_papers`	[.3896,.4046]	[.4659,.4719]	[.4965,.5037]	[.6068,.6118]	[.6864,.6889]	[.7131,.7156]
`wikipedia`	[.3944,.4301]	[.4428,.4505]	[.4657,.4705]	[.6208,.6279]	[.6550,.6569]	[.6754,.6773]

We report two further checks on the $20/20$ separation in §4.1.1, together with a comparison with the source publication's separate *spread* claim.

For the duplicate-count check, GAGLE's split label had been treated as part of an article's identity, so the same article recurred as several "different" human documents. Every one of the $4{,}474$ cross-split copies checked is byte-identical, and collapsing all duplicate groups takes the human population from $\approx 11{,}700$ rows to $5{,}000$ articles, a $2.34\times$ over-count. The $4{,}474$ is the byte-identity verification count, not the number of rows removed; the reduction itself is the $\approx 6{,}700$ excess copies.

1. The population correction behaves as theory predicts. If the pre-correction human intervals were narrow *because of* duplication, removing duplicates should widen them by $\sqrt{N_{\text{pre}}/N_{\text{post}}}$. Measured widening ratios run 1.27–1.61 with mean 1.455 against a predicted 1.47–1.57 (mean 1.529). Moreover, $20/20$ cells widen. The measured mean sits 5% below prediction, the direction expected when the duplicates were byte-identical rather than independent draws.
2. The direction is not the one that models lack long-range planning story predicts. Machine text has *higher* H than human text in every cell. Under the estimator's own semantics, $H \to 1$ is strong persistence, in the limit verbatim repetition higher H is consistent with the repetition and degeneration behaviors documented for autoregressive decoding (Holtzman et al., 2020; Du & Tanaka-Ishii, 2025), not with richer discourse planning. We flag this as an interpretation of the estimator's semantics, not as a measured mechanism; §5.1 states what would be needed to establish it.

A spread comparison worth recording. The source publication additionally reports that human writing has a *narrower* spread of S than model output. On this corpus, the reported variance-ratio direction agrees with this ordering and survives a length control: the variance ratio $\mathrm{Var}_{\mathrm{LLM}}/\mathrm{Var}_{\mathrm{hum}}$ of S exceeds 1 significantly in 8 of 10 (domain, scorer) cells (largest 1.457, `scientific_papers`/Mistral), with the two dissenting cells (`billsum`) null rather than reversed; residualising on log document length moves every ratio by at most 0.016, and the direction holds in $5/5$ length quintiles for both scorers. Bootstrap and Brown–Forsythe agree on direction in every cell. *The same is not true of* H, raw H spread ratios are all significant at 1.19–1.89, but after length residualisation only six remain significant and three reverse, two of them significantly (`wikipedia`/Mistral $1.243 \to 0.813$; `big_patent`/Mistral $1.242 \to 0.883$). An H spread difference should not be reported as a structural property without the length adjustment.

F Full Split-Half Reliability Table

All twenty candidate quantities reached the gate, and none passed. The twelve reported in the source write-up are reproduced below; the remaining eight, including the fit-quality diagnostics S_R^2 and H_R^2, are in `feature_store/p2_reliability_report.json`. R^2 rows are fit-quality diagnostics rather than features, and their low reliability is expected.

Table 2: Split-half reliability, all reported rows. Bar $\rho \geq 0.60$.

Quantity	n	r (half)	ρ (S–B)	Verdict
`gemma_H`	221,904	0.405	0.576	drop (near miss)
`mistral_H`	223,834	0.376	0.547	drop
`gemma_H_window_mean`	221,904	0.275	0.431	drop
`mistral_H_window_mean`	223,834	0.247	0.396	drop
`gemma_H_window_drift`	221,904	0.192	0.322	drop
`mistral_H_window_drift`	223,831	0.167	0.286	drop
`gemma_H_window_var`	221,904	0.162	0.279	drop
`mistral_H_window_var`	223,834	0.158	0.273	drop
`gemma_S`	196,971	0.032	0.063	drop
`mistral_S`	195,552	0.025	0.050	drop
`gemma_S_window_mean`	122,329	0.008	0.017	drop
`mistral_S_window_var`	116,669	0.002	0.003	drop

Method detail. Halves are 200 tokens, so they use the reduced scale sweep $\{8, 16, 32, 48, 64, 96\}$ already validated for length-200 series; the full sweep reaches $\tau = 320$, which is meaningless on 200 points. Windowed quantities use a reduced $100/50$ window because the production $200/100$ window yields exactly one window on a 200-token half, making drift undefined and variance degenerate. They therefore measure the stability-profile *concept*, not the exact production column. Per-column masking with no row dropping, so each row reports its own n. The reduced n on the $\widehat{S}$ rows (196,971 of 223,834, $\approx 120{,}000$ on the windowed variants) is itself informative: on those documents, at least half produced a degenerate Hölder fit.

F.1 Estimator bias against ground truth, full ladder

§4.1.2 summarises this as a document-support bias of -0.26 to -0.48; the full two-point ladder against synthetic fractional Brownian motion of known exponent, run before the corpus was built:

Table 3: $\widehat{S}$ bias against known synthetic ground truth, by support.

Support	$\widehat{S}$ bias vs. truth	fit R^2
One 400-token document, $n_{\mathrm{boot}} = 1000$–5000	-0.26 to -0.48	0.08–0.38
$\approx$80,000 points (200×400 concatenated), $n_{\mathrm{boot}} = 2000$	-0.004 to -0.065	0.91–0.995

Raising the number of bootstrap draws at fixed support does not close that gap. More draws do not fix a short window. In the pooling mode, the corpus-level numbers actually used, bias grows with the true exponent -0.072, -0.160, -0.231 and -0.310 at $H = 0.30$, 0.50, 0.65 and 0.80.

F.2 RELIABILITY UNDER ALTERNATIVE SPLIT CONSTRUCTIONS

The full run behind §4.1.2's split-construction test, eight constructions, all four headline quantities, both the pooled and the stratum-demeaned within-population coefficient (§F.3 defines ρ_{within}).

Whether decimation changes what is measured. Both estimators are functionals of sequential geometry, so taking every other token invites the objection that the halves no longer describe the same object as the whole. Under the premise that makes H and S well defined at all, a process exactly self-similar with stationary increments, they do: decimation by two is precisely the dilation the scaling law is invariant under, and each half carries the same exponent up to a known constant. What the objection reaches is the gap between exact scaling and real text, and that gap is an empirical quantity rather than one an argument can close; the block-size sweep below measures it.

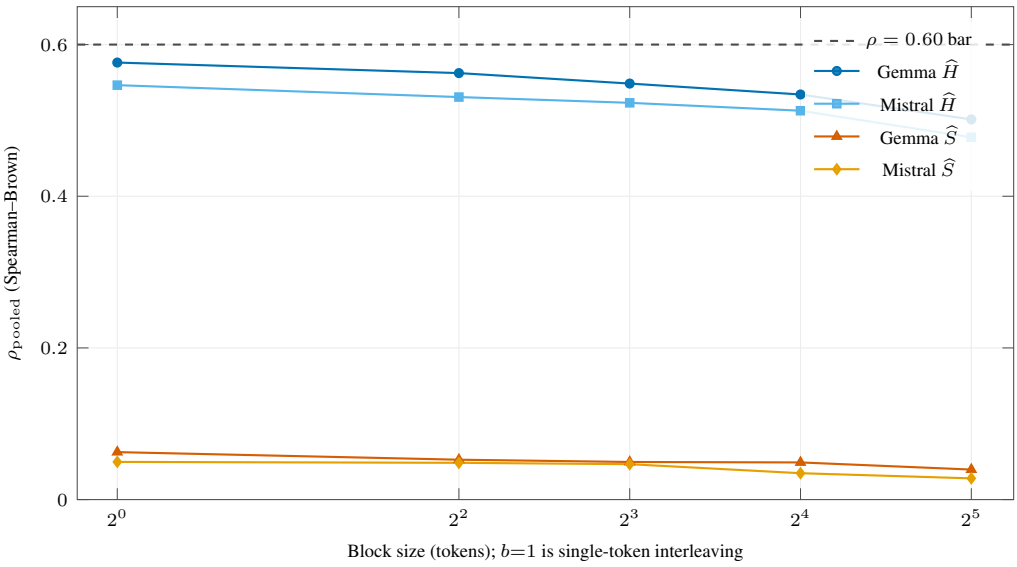

Figure 2: Reliability against block size, $b = 1$ is single-token interleaving. The decline is smooth and monotonic in the direction that argues *against* the sequence-geometry objection, coarser blocks, which decimate the sequence less, are systematically *less* reliable, not more.

Table 4: Split-half reliability (ρ_{pooled}) under all eight tested constructions. `block_b` assigns contiguous blocks of b tokens alternately to the two halves, so `interleave` is the $b = 1$ case and `contiguous` the single-block limit. `block_bootstrap` is measured but structurally invalid here, its two halves share $\approx 47\%$ of token positions by construction (block length ≥ 96 on a ≈ 400-token document), so its inflated correlation reflects token overlap, not document-level signal, and it is excluded from the headline claim. `permutation` is the median of 20 random-membership replicates (order preserved within each half), [p5,p95] in brackets. `t1_permutation_stats.json`.

Construction	gemma_H	gemma_S	mistral_H	mistral_S
interleave (headline)	0.5763	0.0626	0.5465	0.0497
block_4	0.5624	0.0526	0.5308	0.0485
block_8	0.5486	0.0496	0.5232	0.0467
block_16	0.5342	0.0490	0.5128	0.0348
block_32	0.5013	0.0396	0.4777	0.0280
contiguous (pre-registered as confounded)	0.2609	0.0159	0.2678	0.0040
permutation, 20 reps (median [$p5$, $p95$])	0.5443 [.5424, .5462]	0.0530 [.0484, .0571]	0.5152 [.5129, .5182]	0.0452 [.0386, .0498]
matched_sweep (full 6-value `WINDOW_SCALE`)	0.5763	0.0626	0.5465	0.0497
`block_bootstrap` (excluded, see caption)	0.7779†	0.2047†	0.7769†	0.1912†

The headline robustness test. The plan's own pre-registered bar asks whether `gemma_S` stays inside $[0.03, 0.12]$ across every valid construction. It does not, `contiguous` falls to 0.0159. That is the *expected* signature of the drift confound. The interleaved design was built to avoid `gemma_H` collapses in the same construction, $0.576 \rightarrow 0.261$, consistent with genuine within-document stylistic drift entering a contiguous split but not the other constructions. We therefore read it as confirming,

rather than undermining, the case for interleaving. The seven remaining valid constructions (interleave, four block sizes, the permutation median, and the matched full sweep) span 0.0396–0.0626, comfortably inside the band.

An order-scrambling construction, kept as a diagnostic. Among the constructions swept for the permutation check is one built from an *unsorted* random index set for each half, which scrambles a half's internal token order along with its membership. Because both estimators are functionals of ordered structure, this construction collapses ρ to ≈ 0.003 for every quantity, in sharp contrast to every order-preserving construction tested. We keep it here as a direct demonstration that these quantities are genuinely order-dependent: destroying order destroys the signal, which is a magnitude or length statistic computed on the same tokens would not do.

F.3 Variance decomposition (Route 1 and Route 2)

Route 1 demeans both halves by their own $(\text{class}, \text{domain}, \text{temp})$ stratum before correlating the number quoted in §4.1.2 as ρ_{within}. Route 2 is a closed-form, one-way variance-components decomposition (method-of-moments, not an iterative mixed model, chosen for zero convergence risk at $n \approx 223{,}834$) that additionally reports the population share and a cluster-robust standard error, clustered by seed article.

Table 5: Variance decomposition. ICC is the fraction of single-half variance explained by $(\text{class}, \text{domain}, \text{temp})$ stratum membership, i.e. an estimate of $\mathrm{Var}(\mu)/(\mathrm{Var}(\mu)+\mathrm{Var}(\delta)+\mathrm{Var}(\eta))$.

Quantity	ρ_{pooled}	ρ_{within}	ICC	$\mathrm{Var}(\delta)$	cluster-robust SE
`gemma_H`	0.576	0.552	3.9%	0.00217	0.0023
`mistral_H`	0.547	0.520	3.9%	0.00190	0.0024
`gemma_S`	0.063	0.060	0.14%	0.01049	0.0024
`mistral_S`	0.050	0.046	0.18%	0.00870	0.0023

Reading the two exponents side by side, for $\widehat{H}$, population structure (class, domain, temperature together) explains only $\approx 4\%$ of single-half variance, so the $0.576 \to 0.552$ drop is real but modest most of what §4.1.2 measures for $\widehat{H}$ was already within-population. For $\widehat{S}$, population structure explains under 0.2% of variance, $\widehat{S}$'s poor pooled reliability is essentially all within-document estimation noise, ruling out "pooling across populations inflates apparent unreliability" as an explanation for $\widehat{S}$ either way. *Note*, GAGLE has no generator column distinct from `Label` (Human/Gemma-2B/Mistral-7B already is the class–generator join), so the demeaning set is $(\text{Label}, \text{domain}, \text{temp})$ throughout the same information content as $(\text{class}, \text{generator}, \text{domain}, \text{temp})$, without a redundant fourth column. Prompt type is deliberately excluded from the demeaning set throughout this appendix, because §4.2.2 treats what is predictable from the prompt type as part of δ; demeaning by it here would remove the very quantity that section measures.

F.4 Reliability as a function of support, and $N^\star$

The construction behind §4.1.2's $N^\star$ claim. An estimator noise floor $\mathrm{Var}(\eta_N)$ is calibrated against synthetic fractional Gaussian noise (Davies–Harte generator) at $H \in \{0.60, 0.65, 0.70, 0.72\}$, averaged to match GAGLE's own observed $\widehat{H}$ band; $\mathrm{Var}(\delta)$ is taken from Table 5; and $\rho_{\text{SB}}(N) = 2r(N)/(1 + r(N))$ for $r(N) = \mathrm{Var}(\delta)/(\mathrm{Var}(\delta) + \mathrm{Var}(\eta_N))$ is composed at each grid point $N \in \{100, 200, 400, 800, 1600, 3200, 6400, 12800\}$. $N^\star$ is the first grid point at which $\rho_{\text{SB}}(N)$ crosses 0.60; the true crossing may lie between grid points.

What the real-text bracket does and does not establish. A companion pass measures within-document block variance of $\widehat{H}$/$\widehat{S}$ across disjoint N-token blocks on $5{,}000$ full-length human documents (uncapped up to the corpus's own $2{,}047$-token limit). At the matched N values, this real-text variance runs at or below the calibrated fBm floor (e.g. at $N = 400$: $\widehat{S}$ variance 0.0603 real-text against 0.0707 fBm; $\widehat{H}$ variance 0.0046 against 0.0045, materially the same). An earlier draft of this analysis reads that gap as an upper bound on $\mathrm{Var}(\eta_N)$; it is not, blocks drawn from one document share that document's own realization, which suppresses between-block variance relative to

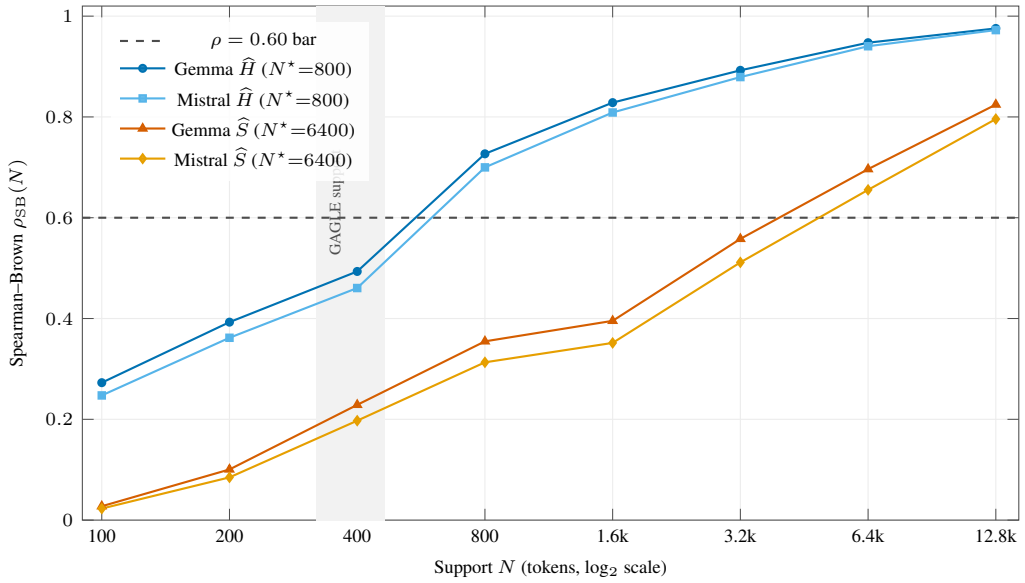

Figure 3: Reliability against support, calibrated fBm noise floor. Shaded band, GAGLE's own 320–464-token usable-support range. $\widehat{H}$ (blue, sky) crosses $\rho = 0.60$ near N=800; $\widehat{S}$ (vermillion, orange) does not cross until N=6,400.

Table 6: $N^\star$, first grid point where the calibrated $\rho_{\mathrm{SB}}(N)$ crosses 0.60. Full curve (all eight grid points) in `review_results/data/t2_n_star_composition.json`.

Quantity	Var(δ)	ρ_{SB} at N=400	$N^\star$
`gemma_H`	0.00217	0.493	800
`mistral_H`	0.00190	0.460	800
`gemma_S`	0.01049	0.229	6,400
`mistral_S`	0.00870	0.197	6,400

the independent draws, the fBm floor uses, an effect distinct from and opposed to the nonstationarity, a bound argument would be needed. What the bracket *does* establish, via a calibration check (median full-document $\widehat{H} = 0.629$ across all 5,000 documents, squarely inside the 0.60–0.72 calibration band), is that the fBm floor is not an underestimate of real-text estimator scatter *for* $\widehat{H}$. The check does not extend to $\widehat{S}$, and there the floor is measurably optimistic, the composed $\rho_{\mathrm{SB}}(400)$ of 0.229/0.197 (Table 6) sits 3.7–4.0× above the 0.063/0.050 measured at the same support (Table 2), so it understates $\mathrm{Var}(\eta_{400})$ for $\widehat{S}$. $N^\star$ is therefore a point estimate under a validated floor for $\widehat{H}$, and a *lower* bound for $\widehat{S}$, correcting the floor moves $\widehat{S}$'s crossing further out, never nearer.

F.5 Estimator invariance, part 1, detrended fluctuation analysis and local Whittle

This appendix ask the same question with three structurally different tools, is the short-support reliability ceiling a property of *what* is being measured, or an artifact of *how* one particular estimator measures it? DFA and the local Whittle spectral estimator are each designed, for different reasons, to be more robust than the plain rescaled-range (R/S) statistic used as this paper's headline to the kind of nonstationarity a short, topically drifting document introduces. DFA fits and subtracts a local polynomial trend from the surprisal walk before measuring its fluctuation at each scale, so slow within-document drift does not get folded into the exponent; the local-Whittle estimator instead works in the frequency domain, fitting the low-frequency slope of the sequence's power spectrum, which is less sensitive to a handful of extreme surprisal spikes than a windowed range statistic is. If

either recovers substantially higher reliability than R/S at the same support, the ceiling would be an R/S artifact rather than a property of the estimand; if not, it generalizes.

Refitting DFA (order-1, same scale sweep) and a local-Whittle spectral estimator on the identical interleaved halves used throughout this appendix.

Table 7: Reliability by estimator. `passes` marks $\rho \geq 0.60$. Local-Whittle saturation (its bounded optimiser hitting $[-0.49, 1.49]$) affects 0.0001 of pairs for both scorers and moves reliability by < 0.001 when excluded, not the source of its clearance.

Quantity	n	ρ_{pooled}	ρ_{within}	Passes 0.60?
`gemma_RS_H` (headline)	221,904	0.576	0.552	no
`gemma_DFA_H`	221,904	0.629	0.601	yes (by 0.029)
`gemma_Whittle_H`	221,904	0.677	0.647	yes
`mistral_RS_H` (headline)	223,834	0.547	0.520	no
`mistral_DFA_H`	223,834	0.599	0.566	no (by 0.0006)
`mistral_Whittle_H`	223,834	0.642	0.605	yes

Table 8: Cross-estimator correlation, document level (mean of halves). Thresholds fixed before this table was produced, > 0.85 reads as "the same quantity, more precisely measured"; < 0.60 reads as "a different quantity".

Pair	Gemma	Mistral
R/S $\sim$ DFA	0.869	0.864
R/S $\sim$ Whittle	0.759	0.751
DFA $\sim$ Whittle	0.808	0.803

R/S and DFA clear the "same quantity" threshold (0.86–0.87), DFA is a less-noisy estimator of the same Hurst exponent, and its clearance of 0.60 is real but marginal and scorer-dependent, which is why §4.1.2 reports it as "at the margin" rather than as unqualified recoverability. R/S and Whittle correlate more weakly (0.75–0.76), so Whittle's more comfortable clearance (0.677/0.642) is partly a different quantity, not purely a cleaner measurement of the same one, and is reported here alongside this correlation rather than as a standalone headline. Neither estimator has a literature-standard analogue for $\widehat{S}$'s displacement-probability construction; $\widehat{S}$'s reliability is unaffected by this appendix and stands exactly where §4.1.2 puts it.

G Extended Detection Results

The label-shuffle control runs the full harness on permuted training labels and lands systematically *below* chance mean AUROC 0.3693 (sd 0.0427) on the random fold, $0/10$ permutations above 0.5, and all eight folds pass a one-sided criterion, because human rows occupy a sparse extreme region of the feature space that a class-weighted booster on random labels assigns the global prior. Injecting leakage raises it monotonically and past the decision boundary at every non-zero level tested ($0\%: 0.4166$; $2\%: 0.5080$; $5\%: 0.5534$; $20\%: 0.8019$), so the control detects what it exists to detect, and its sensitivity floor sits below a 2% contamination rather than above 5%. This catches a class of failure, an index misalignment, or a metric computed against the wrong vector that no confound configuration could, because all five would be wrong together.

Across all forty paired tests (five configurations $\times$ eight folds) DeLong and the paired bootstrap agree in 40 of 40, and Holm–Bonferroni leaves *three* survivors, all on the `billsum` fold and cutting both ways, two where fractal features are significantly worse ($p = 9.4\times10^{-9}$, $p = 3.2\times10^{-5}$) and one where they are significantly better ($p = 4.6\times10^{-5}$), across three configurations. A cross-fold "fractal is significantly harmful" framing does not survive correction; what survives is narrower and more interesting, the effect is statistically real only where the domain shift is largest, and its sign there depends on how missingness is handled. The sole cell favors the full model, `lodo/big_patent`, does so by 0.0001 AUROC, inside that cell's own bootstrap width. The result is not an artifact of the learner, a shallow multilayer perceptron (one hidden layer, 64 units, imputed and scaled, class-balanced) trails the tree ensemble on all eight folds and collapses to TPR@1%FPR

$= 0.0000$ on two of them, where the tree ensemble never does, and the fractal family does not rescue it.

The substitution test, per fold. Isolating a nine-feature log-perplexity group and running `fractal+logppl` (63 features) against `stat_only` (17) on population-matched folds, the fractal arm loses on all eight of eight folds on both AUROC and TPR@1%FPR, $0.9684/0.6460$ against $0.9795/0.7289$ in-distribution, and $0.8191/0.2050$ against $0.8735/0.3171$ on `billsum` LODO. Swapping fractal features *in* for entropy and rank statistics, at comparable combination shape, makes every fold worse.

Table 9: Feature-family ablation, in-distribution (random split, `full`), $n_{\text{test}} = 43{,}413$, human prevalence 2.23%, 2,000-fold stratified bootstrap. Remaining arms and all eight folds, Appendix G.

Arm	#feat	AUROC	TPR@1%FPR	Reading
`fractal_only`	54	0.6629 [.6458, .6782]	0.0641 [.0429, .0914]	real signal, far below either baseline
`stat_only`	17	0.9795	0.7209	incumbent, perplexity, entropy, ranks
`style_only`	14	0.9919	0.7964	incumbent, discourse and surface form
`fractal+stat`	71	0.9785	0.6836	*below* `stat_only` on both
`all`	85	0.9984 [.9976, .9990]	0.9587 [.9489, .9738]	full model
`all_minus_fractal`	31	0.9984	0.9612 [.9360, .9781]	all 54 fractal features deleted

Remaining arms. Three arms omitted from Table 9 for space, `fractal+style` (68 features) reaches AUROC 0.9917 / TPR@1%FPR 0.8171, within noise of `style_only` on AUROC; `all_minus_length` (83) reaches 0.9974/0.9436, so removing length costs 0.0010 AUROC and 0.0151 TPR@1%FPR; and `length_only` (2) reaches 0.8635/0.0477, a partial detector of *human* text and the priced form of the corpus's dominant confound. Four further diagnostic arms, each of which declared a failure condition in advance, on the same random split and `full` configuration, `provenance_only` reaches AUROC 0.5655 / TPR@1%FPR 0.1196 on the six window-length provenance columns, which prices that exclusion (the eleven exit-gate flags are a separate, disjoint set); `shortwindow_only` (2) reaches 0.5290/0.0580, weakly predictive and demonstrably unused; `missingness_only` (85) reaches 0.5169/0.0305, the missingness pattern alone; and `label_shuffle` lands at 0.3693 with $0/10$ permutations above chance.

Table 10: `all` versus `all_minus_fractal` on all eight frozen group-disjoint folds, `full` configuration, one random in-distribution split, five leave-one-domain-out (LODO), and two leave-one-generator-out (LOGO). Each fold is scored on the frozen test split of its held-out stratum, which is why the LOGO n is a fraction of that generator's corpus count rather than all of it. AUROC at six decimals to make the ties legible. `all_minus_fractal` ties or wins on 7 of 8.

Scheme	Held-out fold	n_{test}	`all` (85)	`a-m-f` (31)	Δ	Winner
random		43,413	0.998397	0.998397	$+1.5\times10^{-7}$	tie
lodo	`big_patent`	45,175	0.998571	0.998468	-0.000103	`all`
lodo	`billsum`	45,146	0.937608	0.941725	$+0.004117$	no-fractal
lodo	`newsroom`	44,718	0.992131	0.992077	-0.000054	tie
lodo	`scientific_papers`	43,749	0.994487	0.995077	$+0.000590$	no-fractal
lodo	`wikipedia`	45,046	0.987009	0.987990	$+0.000981$	no-fractal
logo	`gemma_2b`	22,747	0.987404	0.988510	$+0.001106$	no-fractal
logo	`mistral_7b`	21,635	0.986638	0.987624	$+0.000986$	no-fractal

Two patterns in Table 11 are worth naming. First, `fractal_only` is at or near chance on the hardest fold (`billsum`, AUROC 0.5278) and never exceeds 0.6968 anywhere. Second, the incumbent families trade places by domain `style_only` leads on `big_patent` and `scientific_papers`, `stat_only` on `newsroom` the signature of a detector whose components are domain-sensitive in different directions, and part of why the fused model holds up better than any single family.

The `billsum` LODO error analysis. The hardest fold shows the failure mode the whole 85-feature representation has, independent of the fractal family. AUROC falls only to 0.9376 while TPR@1%FPR falls to 0.4931 a tail-ranking failure, not a calibration artifact, bulk ranking is preserved, and the extreme tail is not. An error analysis on this fold (run on a refit whose headline

Table 11: AUROC / TPR@1%FPR, `full` configuration, gradient-boosted tree, all arms on the seven shifted folds.

Scheme	Fold	fractal (54)	stat (17)	style (14)	all (85)	a-m-f (31)	length (2)
lodo	big_patent	.6274/.0278	.9626/.6034	.9907/.7617	.9986/.9632	.9985/.9539	.9734/.2636
lodo	billsum	.5278/.0311	.8732/.3235	.8797/.0612	.9376/.4931	.9417/.5031	.7234/.0177
lodo	newsroom	.6624/.0406	.9858/.8143	.9385/.3570	.9921/.8921	.9921/.8847	.7566/.0077
lodo	scientific_papers	.6968/.0913	.9810/.7705	.9851/.6336	.9945/.8769	.9951/.8918	.9529/.0819
lodo	wikipedia	.5984/.0464	.9667/.7245	.9561/.4518	.9870/.8178	.9880/.8345	.8132/.0359
logo	gemma_2b	.6163/.0531	.9004/.4171	.9786/.6112	.9874/.7954	.9885/.8079	.8506/.0333
logo	mistral_7b	.6567/.0746	.9089/.4114	.9810/.7065	.9866/.7419	.9876/.7555	.8549/.0555

cell was $0.9476/0.5630$, before a later feature-tokenization correction moved it to the values quoted above; the mechanism, not the level, is what carries forward) finds the driver is prompt type by a wide margin. False-negative rate runs 0.677 for chain-of-thought, 0.676 for excerpt continuation and 0.608 for pretrained continuation, against 0.259–0.350 for keyword- and summary-constrained generation a 1.7–$2.6\times$ gap. Freely generated text on this domain's legal-summary register reads more like human writing. The largest distributional shifts into this fold are style features (`Sent_Len_Std` $+0.595$ standardised, KS 0.255; `Bigram_Repetition_Rate` $+0.263$) the family carrying 63.46% of the model's gain.

Deployment-operating-point diagnostics. Beyond the ranking metrics above, three questions bear on deployment: how the detector behaves at very low false-positive budgets, whether its threshold calibration holds under resampling, and how precision changes with the assumed prevalence of LLM-generated text in the traffic being screened. On the `full` configuration random split ($n = 43{,}413$), partial AUROC restricted to FPR$\leq 1\%$ is 0.9595 and to FPR$\leq 5\%$ is 0.9857; TPR at a fixed 5% FPR threshold is 0.9952, and at a much stricter 0.1% FPR threshold falls to 0.6729 (point estimate, 2,000-fold bootstrap). Re-deriving the operating threshold on resampled folds and checking the FPR it actually achieves against its 1% and 5% targets gives 0.93% $[0.41\%, 1.55\%]$ and 4.95% $[3.61\%, 6.40\%]$ respectively both targets covered within CI, so the thresholds in Table 9 are not overfit to one split. Expected calibration error (15 equal-width bins) is 0.00091; the model's predicted probabilities track observed frequencies closely at this operating point.

GAGLE's own corpus is 97.77% LLM-generated (human text is the rare class by construction of the benchmark); a real moderation deployment is typically the opposite shape, with LLM-generated content the minority of traffic. Because precision is prevalence-dependent even when AUROC is not, Table 12 reports PPV at the 1%- and 5%-FPR thresholds across a range of *assumed* LLM prevalence in deployment traffic, not GAGLE's own training prevalence, up to GAGLE's own composition as a reference point.

Table 12: Positive predictive value at fixed FPR thresholds, by assumed LLM-content prevalence in deployment traffic (distinct from GAGLE's own training-corpus composition, bottom row, included as a reference point). Each cell is $\mathrm{PPV} = \mathrm{TPR}\,\pi / \big(\mathrm{TPR}\,\pi + \mathrm{FPR}(1-\pi)\big)$ at the achieved FPRs (0.93% and 4.95%), with π the assumed prevalence. `full` configuration, random split.

Assumed LLM share of traffic	PPV @ 1%FPR	PPV @ 5%FPR
1%	0.510	0.169
5%	0.845	0.514
10%	0.920	0.691
25%	0.972	0.870
50%	0.990	0.953
97.77% (GAGLE's own composition)	0.9998	0.999

H Feature Importance Accounting

Split gain and permutation importance are reported together throughout, because they disagree. Table 13 gives the family totals.

Table 13: Gain versus permutation accounting, `all` arm, random split. Permutation importance is the AUROC drop when a column is shuffled on the test fold, five repeats.

Family	#feat	total split gain	gain share	summed perm. drop	perm. share
style	14	1,296,092.4	63.46%	2.04×10^{-2}	56.7%
stat	17	633,456.7	31.02%	1.54×10^{-2}	42.8%
fractal	54	112,761.3	5.52%	$\mathbf{1.7 \times 10^{-4}}$	0.47%

Per-column detail. Only two fractal columns appear in the top 25 by gain. The highest, `Gemma_H`, ranks 23rd with 0.28% of gain and a permutation drop of $3{\times}10^{-5} \pm 3{\times}10^{-5}$; for comparison the top column overall, `N_Words_Used`, has 39.99% of gain and a drop of $1.28 \times 10^{-2} \pm 2.1 \times 10^{-4}$. The two cross-scorer divergence features the design's explicit attempt to extract signal from *disagreement* between a 2B and a 7B scorer's estimates on the same document rank 60th and 68th of 85.

Distribution within the fractal family. All 54 columns have a permutation drop below 10^{-4}. Thirty-five are zero or negative; shuffling them improves held-out AUROC, and thirty-eight lie within their own standard error of zero. The highest-ranked fractal column by gain is `Gemma_H` at overall rank 23 (0.28% gain, 237 splits, permutation drop $3 \times 10^{-5} \pm 3 \times 10^{-5}$); the lowest is `Mistral_S_window_max` at rank 85. Only two fractal columns appear in the top 25.

Reading the disagreement. Gain counts splits weighted by local objective improvement, which a greedy learner will accumulate on any column with exploitable within-node variation, informative or not. Permutation measures what the fitted model loses on held-out data without the column. For a column that is noise around a population-conditional mean, the first is substantial and the second is nil, which is exactly the pattern observed. The same caution applies in the other direction, and we apply it symmetrically, `N_Words_Used` holds 39.99% of gain, but removing length from the model costs only 0.0010 AUROC and 0.0151 TPR@1%FPR (Appendix G). Gain share overstates the necessity in both cases.

I Multiclass and Auxiliary Results

Crossing corpora costs recall at 1% FPR $0.9587 \to 0.2277$, 95% CI $[0.2127, 0.2391]$ (paired bootstrap, 2,000 resamples) the largest single degradation in this evaluation, quoted in §4.2.3.

Table 14: Cross-corpus transfer, population-matched, both models scored on the identical 22,041-row RAID complete-case population, paired bootstrap, 2,000 resamples. The one place in this evaluation where the fractal family shows a significant advantage.

Model	n	AUROC	95% CI
`all` (85 features)	22,041	0.8032	[0.7975, 0.8089]
`all_minus_fractal` (31, same rows)	22,041	0.7792	[0.7730, 0.7851]
Paired bootstrap ΔAUROC $(85 - 31) = +\mathbf{0.0240}$, 95% CI [0.0222, 0.0259], excludes zero			

Table 15: Auxiliary task results, for scope rather than for the fractal argument. Populations are named per row because n is not comparable across them.

Task	Population	n	Metric	Value
Generator family (Gemma vs. Mistral), 2-class	GAGLE complete-case	14,805	AUROC	0.9943 [0.9936, 0.9951]
same	same	14,805	TPR@1%FPR	0.9016 [0.8871, 0.9131]
Prompt type, 8-class, all features	GAGLE complete-case (85)	14,805	macro-F1	0.5816 [0.5747, 0.5885]
Prompt type, 8-class, `fractal_only`	GAGLE complete-case (54)	14,919	macro-F1	0.1981 [0.1919, 0.2043]
Generator ID, 11-class	RAID complete-case	1,572	macro-F1	0.5441 [0.5164, 0.5684]
Edited vs. clean (single unpaired doc)	edit battery	7,439	AUROC	0.8705 [0.8602, 0.8820]
same	same	7,439	TPR@1%FPR	0.1420 [0.1049, 0.1913]

The same corpus supports the eight-class task at 0.58 macro-F1 with the full feature set and at 0.1981 with the fractal family alone, against a 0.125 chance floor. The two prompt-type arms differ in n by 114 rows because complete-case masking is computed over a different feature count; that cannot

explain a gap of 0.38 macro-F1. The edited-versus-clean row deserves one note, because it was nearly reported at a much stronger level. A first-pass per-variant breakdown at the naive 0.5 threshold showed uniform 93.9–97.8% recall across all six operators and read as a generalized editedness signal. Checking the clean-side false-positive rate at that same threshold first gave 50.3%; half of the clean documents were also called edited, so the table was measuring threshold permissiveness, not discrimination. Recomputed at the clean side's own 5%-FPR operating point, per-operator recall is 37.6–46.5%, still genuinely uniform, so the "not dominated by one operator" conclusion survives, at realistically modest numbers. Detectable in aggregate, weak at low FPR, from four scalar features.

I.1 P3 UNDER POPULATION CONTROL

The full run behind §4.2.2's population-control claim, a population-only control arm, a factorial-balance check on the post-gate population, and group-mean removal on the fractal columns, all against the same frozen prompt-type task (`full`, `lgbm`). The gate reproduces the frozen numbers to within their intervals (`all` 0.5837 against frozen 0.5816 [0.5747, 0.5885], inside the interval but not identical; `fractal_only` 0.1981 exact).

Table 16: P3 population-control results.

Arm	n_{test}	macro-F1	Reading
Population-only (generator+domain+temp, no prompt type)	42,444	0.0594 [.0578, .0609]	*below* 0.125 chance
Fractal-only, before demeaning	14,919	0.1981 [.1919, .2043]	headline number
Fractal-only, after group-mean removal	14,919	0.2046 [.1986, .2111]	rises, does not fall
All-features, after fractal columns demeaned	14,805	0.5831 [.5759, .5899]	unchanged from 0.5816

The population-only control. Given only the generator, the domain, and temperature, no prompt type, no fractal features, no anything else, a classifier reaches macro-F1 0.0594, *below* the 8-class chance floor of 0.125. Population identity alone carries essentially no prompt-type signal, whatever the fractal-only arm's 0.1981 reflects, it is not leaking in through the population factors this control isolates.

Balance table. Across the 30 observed (generator, domain, temp) cells crossed with 8 prompt types (a 240-cell factorial by design), the maximum deviation of any cell's observed proportion from the design's uniform expectation is 0.0022 (0.22 percentage points, $\chi^2 = 413.3$ on 203 d.f.; the associated $p < 10^{-16}$ is a large-n artefact of 14,805 rows, not evidence of a meaningful imbalance, the raw deviation is the number that matters here. The factorial design survives the token-window and complete-case gates essentially intact.

Group-mean removal. Subtracting each of the 54 fractal columns' train-only (generator, domain, temp)-stratum mean, and refitting is the direct test of whether 0.1981 is the residual population leakage; if it were, demeaning should push the number toward chance. It does the opposite, $0.1981 \rightarrow 0.2046$, confidence intervals overlapping, while the all-features arm is unchanged ($0.5837 \rightarrow 0.5831$). Per the plan's own pre-registered interpretation rule (fixed before this table was produced), a result that holds near the undemeaned value means the fractal family carries a genuine document-level prompt-type signal, not population-structure leakage. The number this analysis exposes as most uncertain is 0.5816, the all-features reference, not the fractal-only 0.1981 that the residual-confound objection targeted, had the confound been real, 0.1981 would have fallen, but it did not.

Within-stratum distribution. Refitting separately inside all 30 cells (rather than pooling and controlling) gives all-features macro-F1 mean 0.567 (sd 0.032, range $[0.506, 0.666]$) and fractal-only mean 0.200 (sd 0.026, range $[0.147, 0.255]$), both families sit well clear of chance in every one of the 30 cells, so the pooled 0.1981 is not an artifact of a few cells carrying the signal for the rest.

I.2 RAID GROUP-MEAN REPLACEMENT AND SUBTRACTION

The direct test of §4.2.3's population-mean account of the +0.0240 RAID transfer gain, run on the identical 22,041-row complete-case population as Table 14 (baseline AUROC reproduces exactly, 0.80321).

Table 17: RAID group-mean replacement and subtraction. $(\text{generator}, \text{domain})$ groups, the eight domains crossed with eleven generators and the human pool, 96 groups total.

Arm	AUROC	% of $+0.0240$ gap	Reading
`all` (85, reference)	0.8032	100%	baseline
`all_minus_fractal` (31, reference)	0.7792	0%	baseline
Replacement, fractal cols. $\to$ group mean	0.8149	149%	group mean alone *exceeds* the real gap
Subtraction, fractal cols. $\to x - \bar{x}_k + \bar{x}$ (k the group)	0.7934	59%	part of the gap removed (41%)

Replacing each document's 54 fractal values with its own $(\text{generator}, \text{domain})$ group mean does not merely recover the gap, it *exceeds* it, group-mean-only fractal columns outperform the real per-document values on this transfer task. Subtracting the group mean (while adding back the global mean, so the frozen model still sees in-range inputs rather than values centered near zero, it never trained on) retains 59% of the gap rather than collapsing it to zero. Read together, population-level $(\text{generator}, \text{domain})$ information does the majority of the work in the RAID transfer gain, with a real, non-trivial document-level residual left over tilted toward the population-mean account, but not a clean "it is entirely the population means story.

J Perturbation Battery, Full Detector Comparison and Arm Decomposition

Table 18: Recall@1%FPR with and without the fractal family, by operator class. *The class predicts the sign of the contribution in all six cases.* Row counts differ slightly because the 85-feature numbers use the Binoculars-limited $39{,}773$-row intersection and the 31-feature numbers the full $39{,}810$-row edited store.

Class	Operator	n (85)	n (31)	`all` (85)	`a-m-f` (31)	Contribution of fractal family
–	(pre-attack, clean)†	42,444	42,444	0.9559	0.9597	-0.38 pt (tie; see caveat)
II	`paraphrase_light`	5,157	5,162	0.9639	0.9545	$+0.94$ pt
	`paraphrase_heavy`	5,255	5,265	0.9243	0.9033	$+2.10$ pt
	`style_transfer`	5,380	5,393	0.9520	0.9425	$+0.95$ pt
	`summarize_then_expand`	4,293	4,298	0.9846	0.9744	$+1.02$ pt
I	`sentence_reorder`	9,831	9,835	0.2670	0.2866	$\mathbf{-1.96}$ pt
	`typo_noise`	9,857	9,857	0.5851	0.6727	$\mathbf{-8.76}$ pt

Why a sign flip is the predicted outcome. Under $\mathrm{Var}(\eta) \gg \mathrm{Var}(\delta)$ a large structured perturbation enters an already noise-dominated estimate, the learned thresholds read the wrong side, and a model that never held the quantity is better off than one holding a corrupted version of it. `sentence_reorder` preserves every token, every word frequency and every lexical statistic while permuting the order in which they arrive; `typo_noise` preserves the content and perturbs the tokenization. Everything else in the model is a content statistic, and content is exactly what these two operators preserve, which is why they, and only they, invert the family's contribution.

Operator definitions. *Class I (index structure).* `sentence_reorder` permutes contiguous blocks, destroying the ordering that the rescaled range integrates over; `typo_noise` performs character-level swaps, drops and slips that re-tokenize the affected words, shifting every downstream index and injecting high-surprisal spikes into the increment sequence. *Class II (realization).* `paraphrase_light`, `paraphrase_heavy`, `style_transfer` and `summarize_then_expand` rewrite vocabulary and register at fixed coarse-scale ordering.

The two qualifications on §4.3. An arm decomposition localises the Class I *collapse* in the non-fractal families, `fractal_only` being already at 5.8% pre-attack and barely moving. And the pre-attack row † of Table 18 is a copy-through from an operating-threshold file we could not re-derive, whereas all twelve per-attack recall values (six operators $\times$ two arms) reproduce *bitwise* from the stored thresholds; re-deriving them narrows `typo_noise`'s contribution from -8.76 to roughly -7 points without approaching zero.

The operator partition generalizes beyond this estimator family, for any representation defined as a functional of some structure in the input, an operator acting on that structure and one acting around it constitute a two-sided test of whether the representation encodes the structure or merely correlates with the population that produced it.

Table 19: Recall@1%FPR by detector and operator, shared population $n = 39{,}773$. ↑ marks recall *above* the detector's own pre-attack baseline.

	Pre-attack		Class II, realisation				Class I, index	
Detector	@1%	@5%	para. light	para. heavy	style	summ.→exp.	reorder	typo
`roberta`	0.9989	0.9999	0.9996	0.9989	0.9993	0.9991	0.9959	0.9968
`tabular` (85)	0.9559	0.9906	0.9639	0.9243	0.9520	0.9846	0.2670	0.5851
`binoculars`	0.8728	0.9231	0.8689	0.7922	0.8554	0.7456	0.7846	0.8110
`fastdetectgpt`	0.5104	0.6482	0.6735↑	0.6344↑	0.6898↑	0.8600↑	0.2430	0.1967
`perplexity_gemma` *(floor)*	0.0935	0.3339	0.0076	0.0076	0.0063	0.0005	0.0183	0.0041

One caveat is stated once, rather than per row, a detector whose *pre-attack* recall is already near the floor (`perplexity_gemma`, 0.0935) has no signal for an attack to destroy or spare, and its post-attack numbers describe noise around an uninformative baseline, not robustness.

Arm decomposition, in full. Refitting single-family arms on the detection stage's own training rows (`config=full`, random/random) and scoring them against `sentence_reorder`: `all_minus_fractal` (31 features, stat and style pooled) falls 94.3% → 22.7% recall@1%FPR, a 71.6-point drop; `stat_only` (17) falls 80.6% → 34.5% (46.1 points); `style_only` (14) falls 77.6% → 27.6% (50.0 points); and `fractal_only` (54) was already at 5.8%/17.4% (at the 1% and 5% operating points) pre-attack and barely moves under attack (3.2%/14.2%) a floor effect, not evidence of robustness. Two qualifications, because the arithmetic invites an overclaim in each direction. (i) The drops are *not* additive, at 5% FPR the pooled arm degrades *less* than `style_only` alone (39.0 against 44.0 points), so pooling gives the model a redundant signal to fall back on at that operating point rather than simply summing two overlapping weaknesses, which is why we report post-attack absolute recall alongside drop size rather than asserting an additivity pattern that the numbers do not support. (ii) These arms are refits on the detection stage's own rows, a different population from the 39,773-row shared intersection of Table 19, so they are read for direction and relative magnitude, not merged with that table's levels.

K Attacker Capability Ladder

A note on floors first, since §4.4 relies on it, we treat a detector whose *pre-attack* recall@1%FPR is already at or below 15% as having no usable signal for an attack to destroy or spare, so its post-attack movement describes noise rather than robustness. Only `perplexity_gemma` (9.35%) falls below that screen; Fast-DetectGPT's 51.04% baseline sits far above it.

Table 20 gives the sweep behind §4.4, a single operator (paraphrase) regenerated by a size sweep of open instruction-tuned models on a 5,000-document population, with a cross-family control. A sixth candidate (Falcon-7B-Instruct) was dropped for instruction-following failure rather than for its results.

Table 20: Recall@1%FPR under paraphrase by attacker model. The first four columns are the size-ordered Qwen2.5 sweep; Phi-3.5-mini is a cross-family control and is not part of that ordering, so the pattern column describes the Qwen sweep only.

Detector	Qwen2.5 0.5B	1.5B	7B	14B	Phi-3.5-mini (3.8B)	Pattern
`tabular`	72.6%	43.5%	45.1%	36.5%	37.7%	broadly declining across the Qwen sweep
`binoculars`	70.8%	58.6%	59.8%	35.1%	45.3%	cleanest decline; one inversion at 7B
`roberta`	99.7%	99.8%	98.8%	99.4%	99.9%	unmoved at every attacker size
`fastdetectgpt`	3.2%	2.4%	4.5%	1.2%	1.7%	floor-level under any non-lineage attacker

Regenerating the same paraphrase with models outside the scorer's lineage reverses the Class II effect entirely. Fast-DetectGPT's recall moves from 0.63–0.86 under the same-lineage paraphrase to 0.012–0.045 across the five attackers above, a reversal large enough that no intermediate read-

ing is ambiguous. One clarification on provenance, an earlier two-model diagnostic (Qwen2.5-7B-Instruct, Phi-3.5-mini-instruct), run before this ladder and on a separate population, put the same reversal at 0.025–0.157. It is the prerequisite investigation that prompted the ladder, not an independent measurement of it; the ladder supersedes it, and the range quoted in §4.4 and here is the ladder. Reporting the attacker's lineage alongside the attack type would make the exposure visible without running anything.