MARY HAD A LITTLE [WHAT?]: DECODING CAN MATTER MORE THAN THE MODEL ITSELF IN ZERO-SHOT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Likelihood-based zero-shot detectors seek to identify the source of a text by comparing its observed tokens with what an evaluator LLM expects to see. We find that this relationship can depend less on who wrote the text than on how it was sampled, which is a routine generation choice. Across six evaluator LLMs and eight domains, applying standard truncation (top-k/top-p) to an evaluator's own generation moves its liketropy score about twice as far as switching to another LLM, and more than ten times as far as switching to a human writer. Calibrated to accept 95% of the evaluator's raw-sampled text, liketropy rejects 78% of its truncated text yet accepts 88% of human text as its own. To explain this, we derive a decomposition of the expected score for a family of log-likelihood-based detectors. It combines an entropy-based configuration term, which is highly sensitive to how narrowly the generator samples, and an identity term, which depends on the KL divergence from the generator to the evaluator. Consistent with this mechanism, in a complementary experiment involving five detectors, switching the generator from truncated to full-vocabulary sampling drags AUROCs as high as 0.99 down to chance or below, while a supervised classifier outside the family still separates the same texts with AUROC 0.84.

1 INTRODUCTION

Determining whether a text was written by a human or produced by a particular large language model (LLM) has become important across academia, social media, and online commerce (Sadasivan et al., 2025; Gambetti & Han, 2023; Salminen et al., 2022). Supervised detectors learn textual regularities from labeled human- and machine-generated examples (Jindal & Liu, 2008; Ott et al., 2011; Bakhtin et al., 2019; Solaiman et al., 2019; Guo et al., 2023; Liang et al., 2023), but their performance can deteriorate across domains and generator families, and adapting them to new models may require additional training and labeled data. Zero-shot likelihood-based methods offer an appealing alternative. Without training a separate detector, they evaluate the observed tokens using probabilities, ranks (Gehrmann et al., 2019), entropies, or perplexities supplied by one or more evaluator language models. Examples include the log-likelihood (Solaiman et al., 2019), the mean token entropy (Mitchell et al., 2023), GLTR (Gehrmann et al., 2019), DetectGPT (Mitchell et al., 2023) and its analytic successor Fast-DetectGPT (Bao et al., 2024), Binoculars (Hans et al., 2024), and the liketropy test of Radvand et al. (2026). These methods share a central idea: text produced by a model should exhibit a recognizable statistical relationship with what an evaluator language model expected to observe.

This paper explores the limits of such likelihood-based zero-shot detection methods. We study these limits using *liketropy* as our primary case study, which is particularly informative for this purpose because it is formulated as a statistical test for model attribution and is accompanied by concentration results and finite-sample error bounds (Radvand et al., 2026). We start with a surprising empirical observation in Section 2. For a fixed LLM evaluator, the liketropy gap between the model's own output and human text, or even text from an entirely different LLM, is substantially smaller than that caused by simply truncating that same model's sampling distribution, which is a routine decoding choice that is often treated as an implementation detail.

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.

Section 3 examines the underlying mechanisms. We introduce a family of log-likelihood-based methods and show that their expected per-token scores separate into two components. The first is an entropy-based component, which we call the *configuration term*. It is independent of the comparison between the generator and the reference distribution, and is particularly affected by how narrowly the generator samples. The second is a weighted KL divergence from the generator's conditional distribution to the detector's reference distribution, which we call the *identity term*. It is the term that reflects intrinsic differences between the models. However, a zero-shot detector observes only the combination of the two. Separating them requires the generator's distribution, which is available only when the text comes from an LLM that can be queried, and never when it comes from a human. This decomposition reveals two failure modes. The first is *over-sensitivity to truncation*: truncation reduces the entropy of the LLM's output distribution and moves the configuration term far from zero. The detector can therefore reject text produced by the very model being tested. The second is *insufficient sensitivity to model differences*. When the generator and the evaluator have similar entropy, the configuration term is near zero. Meanwhile, the identity term can be too weak to support reliable detection. That leads the detector to accept a foreign model. Our controlled experiments in Section 4 further test and isolate these mechanisms.

Overall, this paper sheds light on the fundamental limits of zero-shot likelihood-based detectors. More broadly, this paper contributes to the limited literature on how decoding choices affect the detectability of LLM-generated text. Most detection studies evaluate a fixed generation pipeline and treat decoding as an implementation detail, with notable exceptions such as Ippolito et al. (2020); Dugan et al. (2024); Kempton et al. (2025). We complement these studies by focusing on the phenomena in zero-shot detection, and identifying the structural mechanism of the effect of decoding on likelihood-based scores.

2 The Phenomenon

Preliminaries. Let us first recap the liketropy framework by Radvand et al. (2026, v5), which we use throughout the paper as a representative case study. An autoregressive LLM produces tokens Y_n from a vocabulary χ of size K. For an evaluator *Model A*, the log-perplexity of an observed string, $l_A(Y_N) = -\frac{1}{N}\sum_{n=1}^{N} \log p_n^A(y_n^{\text{obs}})$, converges to the average entropy $h_N(A,A) = \frac{1}{N}\sum_n H(p_n^A)$ when *Model A* itself generated the text by sampling from its full distribution (their Theorem 1(a)), and to the cross-entropy $h_N(B,A) = \frac{1}{N}\sum_n H(p_n^B, p_n^A)$ when a different *Model B* did. The statistic is thus the difference of the two:

$$\text{liketropy} = h_N(A,A)(Y_N) - l_A(Y_N), \tag{1}$$

which converges to zero if the text is sampled from p^A and to $h_N(A,A) - h_N(B,A)$ if it is sampled from p^B. As a result, a statistical test can be conducted around the hypothesis H_0: *some source* $B \neq A$ *generated the text* versus H_1: *Model A did.* For readability we say the test *accepts* a text as *Model A*'s own when it rejects H_0. Specifically, Radvand et al. (2026) propose two rules. The main rule rejects H_0 when $|\text{liketropy}| \leq t$ for a threshold t near zero. A one-sided variant drops the absolute value and rejects when liketropy $\geq -t$, so its rejection region is the *high* side of the statistic; the drop is based on the assumption that $h_N(B,A) > h_N(A,A)$ for every non-A generator. Radvand et al. (2026) then provide guarantees of error probabilities of this test in the large-deviation regime.

Meaning of the sign. Liketropy compares how surprised *Model A* is by the text with how surprised it expected to be. Two things move it, in opposite directions.

- **Zero** is the reference point: *Model A* wrote the text by sampling from its full distribution, the case Theorem 1(a) is stated for.
- **Negative.** The writer picks words *Model A* would not have picked, so *Model A* is more surprised than expected. This push depends on who wrote the text, and it is the one the test relies on.
- **Positive.** The writer samples more narrowly than *Model A* does, for instance under truncated decoding, and keeps to its own most likely words. These are mostly common words, which *Model A* also finds unsurprising, so *Model A* is less surprised than expected. This push depends on how the text was sampled, not on who wrote it.

A text can be pushed both ways at once, so its sign shows which push is larger. Section 3 derives the two pushes as the two terms of the expected score. Signed values are used throughout.

The main experiment. The liketropy test takes two inputs: an evaluator (*Model A*), the LLM that computes the score, and a text. It asks whether the evaluator generated the text. Every text has an underlying generator, whose identity we know.

The design isolates the generator: the evaluator's settings are held fixed: *Model A* always scores with its full conditional distribution at every position. That is forced rather than chosen: liketropy reads $\log p_n^A(y_n)$ at the token actually observed, and an evaluator restricted to its own top-k set assigns probability zero outside that set, so a single observed token outside that set, whether from a human writer or another model, sends the log-perplexity to infinity. Scoring with the full vocabulary is the only setting in which the statistic is defined, so any failure observed below is internal to the framework rather than a product of our setup.

Six base LLMs serve as both generator and evaluator (Falcon-RW-1B, OPT-1.3B, Phi-2, Gemma-7B, Llama-3-8B and Mistral-7B-v0.1). Spanning five organisations, four tokenizer families and 1B-8B parameters, texts come from XSum (Narayan et al., 2018) and seven RAID domains (Dugan et al., 2024). Only the generator varies, along two axes. The identity axis runs from *Model A* itself to some other source $B \neq A$, a different model or a human writer; human text is the limiting case, a writer the evaluator was trained to imitate. The configuration axis runs from raw full-vocabulary sampling to the truncation the original pipeline configures, top-$k = 40$ and top-$p = 0.96$. Crossing the two gives a 2×2 factorial design: A and A'' (*Model A* wrote the text) and B and B'' (a different model did). The double prime marks the original setup throughout: truncated generation, scored with the first 30 prompt tokens inside the window. Elsewhere the prompt is excluded, leaving a scored window of 200 tokens that are entirely model-generated or entirely human-written. Two further conditions each extend one axis: A_k sweeps the truncation depth, and V replaces the generator with a relative of *Model A*; Table 1 states all seven. Appendix A gives the length control, the prompt treatment and the numerical precision. Table 3 reports the complementary design, which holds one scorer and one set of texts fixed and varies the detector.

Table 1: **Summary of generator and sampling settings.** A cell is one evaluator on one dataset and 400 documents (except under A_k, where it is one evaluator at one truncation depth and 200 documents). Cells = evaluators $\times$ generators $\times$ datasets. †XSum only; every other condition spans all eight datasets.

Condition	Generator	Sampling	Cells	# Texts
A	$=$ evaluator	full vocabulary	$6 \times 8 = 48$	19,200
A''	$=$ evaluator	top-k 40, top-p 0.96	$6 \times 8 = 48$	19,200
A_k	$=$ evaluator	$k \in \{5, \ldots, 640\}$, top-p 1.0	$6 \times 9 = 54^\dagger$	10,800
Human	a human writer	n/a	$6 \times 8 = 48$	19,200
B	$\neq$ evaluator	full vocabulary	$6 \times 5 \times 8 = 240$	96,000
B''	$\neq$ evaluator	top-k 40, top-p 0.96	$6 \times 5 \times 8 = 240$	96,000
V	a relative of the evaluator	full vocabulary	$6^\dagger$	2,400

Observations and discussion. Two of the five rows in Figure 1, A and A'', are texts the evaluator wrote itself; the other three are not. The test should accept the first two and reject the other three. However, A'', which the model did write and which differs from A only in the sampling method, lies furthest from zero, although by equation (1) it should score zero. That is the failure Section 4.1 takes up: the test reports a difference in decoding, not in authorship. At the same time, the test accepts foreign text (human text, B and B'') before it accepts A'', the model's own text generated with truncated sampling. That is the failure Section 4.2 takes up: the test misses a difference that is there. Strictly speaking, truncated output comes from a different distribution than *Model A*'s full one, so rejecting A'' is what the test's theorem allows. We still count A'' as *Model A*'s text, following a common convention in the detection literature: the decoding method is treated as a setting of the model, not as a different writer. The evaluation in Radvand et al. (2026) follows the same convention, labelling truncated generations as *Model A*'s own (Appendix B). Under either reading, the test still accepts much of the text that *Model A* did not write. No choice of cut-off repairs this. Ordered

by distance from zero, the rows are A at 0.026, human text at 0.067, B'' at 0.270, B at 0.285 and A'' at 0.475, so the two conditions the model wrote are the nearest and the furthest, and the three it did not write lie between them. A cut-off wide enough to keep A'' keeps everything else with it. The sign splits the rows by decoding instead: A, B and human text sit below zero, A'' and B'' above it, which is the line between restricted and unrestricted writing rather than the line between the evaluator and the rest. Replication under the framework's own settings reproduces its reported operating point: AUROC 0.987 on XSum for Llama-3-8B against the 0.99 reported by Radvand et al. (2026, App. F.2, Table 5). Moving to full-vocabulary sampling drops it to between 0.359 and 0.555 across the six evaluators (Appendix I). Decoding appears nowhere in the framework's account of what the statistic measures.

Table 3 applies the same manipulation to four further detectors, and Section 3 shows that the pattern is not specific to liketropy.

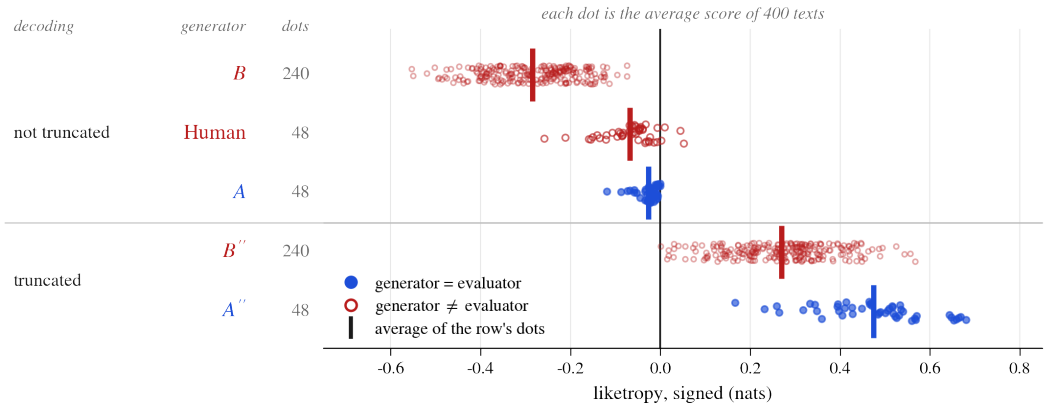


Figure 1: **Five kinds of text, scored by the liketropy test.** Each row is one generator condition. The columns on the left say who wrote the text and whether the writer was restricted while writing, which for a model means truncated decoding at top-k 40 and top-p 0.96. Each dot is one evaluator's mean score over 400 texts from one generator on one of the eight datasets; the bar is the row mean. The B and B'' rows carry five times as many dots because five other models can be the writer: when Llama-3-8B is the evaluator it reads Gemma-7B, Mistral-7B-v0.1, Phi-2, OPT-1.3B and Falcon-RW-1B in turn, one dot per generator and dataset. The test accepts a text as the evaluator's own when the score sits near zero.

3 Understanding the Mechanism

The log-likelihood family of detectors. Let p_n^B denote the conditional distribution from which the generator draws the token at position n, where B ranges over every source the detector may face: another model, a human writer, or *Model A* itself under a decoding configuration other than full-vocabulary sampling. The evaluator is *Model A* with full conditional distribution p_n^A, and $y_{<n}$ is the prefix. The detectors in the log-likelihood family differ in what they compare the observed token against, but they agree on the shape of the comparison.

Definition 1 (The log-likelihood family)**.** *A detector belongs to the* log-likelihood family *if its score on $y_{1:N}$ can be written as*

$$S(y_{1:N}) \;=\; \frac{1}{N}\sum_{n=1}^{N}\big[\,\alpha_n \log q_n(y_n) + \beta_n\,\big], \tag{2}$$

where the reference distribution q_n and the coefficients α_n and β_n are all fixed by the prefix $y_{<n}$ and by the evaluator models the detector holds, not by the token y_n actually observed in the given text.

Detectors that divide one sum by another belong to the family too, since a detector only ever compares its score with a threshold t. For any denominator whose terms $c_n > 0$ are fixed by the prefix,

$$\frac{\sum_n \alpha_n \log q_n(y_n)}{\sum_n c_n} \;\geq\; t \quad\Longleftrightarrow\quad \frac{1}{N}\sum_{n=1}^{N}\big[\alpha_n \log q_n(y_n) - t\, c_n\big] \;\geq\; 0, \tag{3}$$

which is equation (2) with $\beta_n = -t\, c_n$. Binoculars is the case $c_n = H(p_n^{M_1}, p_n^{M_2})$, the cross-entropy between its observer model M_1 and its performer model M_2.

The reference q_n is the probability distribution over all potential tokens that *Model A* assigns to the nth position of the given text: *Model A* reads the prefix $y_{<n}$ and spreads its probability over the vocabulary, and $q_n(y_n)$ is how much of that probability went to the token that actually appeared in the nth position. On average, $-\log q_n(y_n)$ carries two pieces of information at once: the generator's own entropy $H(p_n^B)$ and the KL divergence from the generator to *Model A*. Because p_n^B, the distribution of the generator of the text, is never disclosed to the detector, the observed token is the only place where the writer's distribution meets the reference, and $\log q_n(y_n)$ is the only term that reads it there.

The most common offset β_n is *Model A*'s own entropy, $H(p_n^A) = -\sum_{y\in\chi} p_n^A(y)\log p_n^A(y)$, which is produced the same way as q_n: by *Model A* reading the prefix. Both therefore move whenever the text moves. What the detectors in the family do not account for is that the text changes not only when the writer changes, but also when the sampling method used in generation changes, and Section 2 showed that the second has the larger effect. To find out why, we split the expected score of equation (2) into two parts, a configuration term and an identity term.

Proposition 1 (What a family member reports)**.** *If $y_n \sim p_n^B$ for each n, then any S of the form* (2) *satisfies*

$$\mathbb{E}[S] \;=\; \frac{1}{N}\sum_{n=1}^{N}\Big\{\underbrace{\beta_n - \alpha_n H(p_n^B)}_{\textit{configuration}} \;-\; \alpha_n \underbrace{D_{\mathrm{KL}}(p_n^B \,\|\, q_n)}_{\textit{identity}}\Big\}. \tag{4}$$

Table 2: **Five detectors in the log-likelihood family.** Each row gives the detector's q_n, α_n and β_n in equation (2), which is all it computes; the Binoculars row uses the form of equation (3). The last two columns are the two terms of Proposition 1 at position n, and $\mathbb{E}[S]$ is their sum averaged over the N positions; the detector cannot compute either, since both depend on p^B. Here $H_A = H(p_n^A)$, $H_B = H(p_n^B)$ and $H_{12} = H(p_n^{M_1}, p_n^{M_2})$. GLTR (Gehrmann et al., 2019) has no row of its own. Its token ranks are not affine in $\log q_n$, and its entropy feature, averaged over positions, is the same statistic as the entropy detector in the second row.

				$\mathbb{E}[S]$	
Detector	**Reference** q_n	α_n	β_n	**configuration**	**identity**
log-likelihood	p_n^A	1	0	$-H_B$	$-D_{\mathrm{KL}}(p_n^B \| p_n^A)$
entropy	–	0	H_A	H_A	none
liketropy	p_n^A	1	H_A	$H_A - H_B$	$-D_{\mathrm{KL}}(p_n^B \| p_n^A)$
Fast-DetectGPT	p_n^A	1	H_A	$H_A - H_B$	$-D_{\mathrm{KL}}(p_n^B \| p_n^A)$
Binoculars	$p_n^{M_1}$	-1	$-t\, H_{12}$	$H_B - t\, H_{12}$	$+D_{\mathrm{KL}}(p_n^B \| p_n^{M_1})$

The two terms of equation (4), proved in Appendix C, respond to different things. The identity term is the divergence from the generator's distribution to the reference, zero only when $q_n = p_n^B$. The configuration term combines the offset β_n, set by the evaluator, with the generator's own entropy $H(p_n^B)$ scaled by α_n. Both parts are entropies rather than comparisons between two distributions, which is why we name the term after the decoding configuration: restricting a model to its 40 most likely tokens barely changes which tokens it prefers, but makes its choices far more predictable, moving the identity term by 0.100 nats and the configuration term by 0.545 nats. Every member with $\alpha_n \neq 0$ therefore carries the generator's entropy, and *no member of the family reports the two terms separately.* Liketropy is the case $\alpha_n = 1$, $q_n = p_n^A$, $\beta_n = H(p_n^A)$, which puts the origin on

the evaluator's own entropy and makes the configuration term a difference of entropies:

$$\mathbb{E}_{p_n^B}[\text{liketropy}_n] = H(p_n^A) - H(p_n^B, p_n^A) = \underbrace{\left[H(p_n^A) - H(p_n^B)\right]}_{D_n,\ \text{configuration term}} - \underbrace{D_{\mathrm{KL}}(p_n^B \,\|\, p_n^A)}_{\text{identity term}}. \tag{5}$$

The decomposition already appears in Radvand et al. (2026), but there the entropy difference is folded into an error term rather than treated as one of the two things the statistic measures.

Table 3: **Five detectors, the same manipulation.** Each row is one detector, given the same job twice: look at a piece of text and say whether a model or a person wrote it. On the left the model was restricted while writing; on the right it was not. Nothing else differs between the two runs: the scorer is Gemma-7B for every detector except Binoculars, which uses its own Falcon-7B pair, and each result pools eight datasets. A bar's length is its AUROC: a full slot means always right, and the halfway mark means chance. The rows are grouped by the coefficient α_n of Table 2.

	Truncated LLM vs Human LLM text, top-k 40		**Raw LLM vs Human** LLM text, raw sampling	
Detector		**AUROC**		**AUROC**
carries an identity term ($\alpha_n \neq 0$)				
log-likelihood		0.88		0.28
Binoculars		0.95		0.34
liketropy		0.99		0.52
Fast-DetectGPT		0.99		0.53
no identity term ($\alpha_n = 0$)				
entropy		0.45		0.76

The left column is the setting published evaluations use, and the source of their high scores.

Experiments beyond liketropy. Table 3 applies the same manipulation to four further detectors. Every row is asked one question, the same one liketropy asks: separate Gemma-7B's own text from human text. Under truncation the four detectors carrying an identity term score 0.88 to 0.99; at full vocabulary they score 0.28 to 0.53. The two halves of the table are the two failures side by side. On the left the detectors look accurate, but what they are separating is restricted writing from unrestricted writing rather than one writer from another, which is the failure of Section 4.1. On the right the restriction is lifted and nothing else changes: the text is still the model's own, yet four of the five detectors can no longer tell it from human writing, which is the failure of Section 4.2. Entropy is the only detector that improves, rising from 0.445 to 0.764, and it is the only row with no identity term (Table 2). It reads only how unsure the evaluator is at each position. Full-vocabulary sampling sometimes picks unlikely tokens, and these tend to lead the text into places where the evaluator is uncertain, while human writing does not drift this way. Entropy therefore separates the two writers under raw sampling, but not under truncation. Although every member of the family fails under one of the two decoding settings, the texts themselves stay detectable: two supervised detectors that read surface features score between 0.74 and 0.88 under both generation settings (Appendix J).

4 Testing the Mechanism via Two Failure Modes

Since the detector observes only the sum of the (unknown) configuration and identity terms, two failure modes follow. The first is *over-sensitivity to truncation*: when *Model A* writes under truncated decoding, the configuration term D carries the statistic away from zero, and the model's own text is rejected. The second is *insufficient sensitivity to model differences*: when the decoding is held fixed and the writer changes, only the identity term registers the change, but it is too small to move the statistic away from zero, so the foreign text is accepted. In this section, we design controlled experiments to test these two failure modes separately.

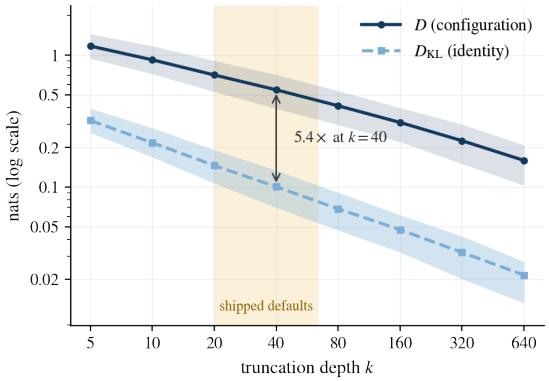
4.1 Failure Mode 1: Over-Sensitivity to Truncation

Let *Model A* write with top-k (Fan et al., 2018) or top-p sampling (Holtzman et al., 2020), which is how most deployments run a model. The text is the model's own, so the test should accept it. Both

sampling methods do the same thing. They keep the set S of most probable tokens, which together hold probability $Z = \sum_{y \in S} p_n^A(y)$, drop the rest of the vocabulary, and divide what is left by Z to renormalise it, written $\widetilde{p}_n^A$. Every kept token is divided by the same number, so the identity term is fixed by that number alone:

$$D_{\mathrm{KL}}(\widetilde{p}_n^A \,\|\, p_n^A) = \sum_{y \in S} \frac{p_n^A(y)}{Z} \log \frac{p_n^A(y)/Z}{p_n^A(y)} = -\log Z, \tag{6}$$

while the configuration term becomes $H(p_n^A) - H(\widetilde{p}_n^A)$, the entropy the cut removed.[1] The two terms now behave differently. The identity term counts only the probability that was cut. The configuration term counts the entropy that was cut, which is much larger, and it grows with every token removed. Nothing holds it down. The score is the difference of the two terms, so it follows the configuration term. The test then reports that the text is far from *Model A* when *Model A* actually generated it. What moved the score was the decoding, and the test has no way to tell decoding from authorship.



(a) both terms as k grows

(b) the same averages numerically

k	D	D_{KL}	D/D_{KL}	liketropy
5	1.170	0.3195	3.66	0.852
10	0.920	0.2165	4.26	0.703
20	0.707	0.1457	4.86	0.564
40	0.545	0.1005	**5.45**	**0.448**
80	0.413	0.0684	6.06	0.344
160	0.309	0.0475	6.52	0.263
320	0.224	0.0319	7.04	0.190
640	0.159	0.0215	7.39	0.133
raw	0.001	0.0017	–	−0.007

Figure 2: **The configuration term is** 3.7 **to** 7.4 **times the identity term at every truncation depth.** Each point averages six models writing XSum text while choosing each word from only their k most likely candidates, with top-p switched off so that k is the only thing that changes; bands span the six models, and units are nats. Both axes are logarithmic, so a constant ratio would show as parallel lines; instead the gap widens as k grows. The shaded band marks the values of k that common software uses by default. In (b), liketropy is the mean score observed on the sampled text. Raw sampling is left out of (a) because both terms are zero in expectation and cannot be drawn on log axes.

Because the writer here is *Model A* itself, equation (6) gives the identity term exactly, so both terms can be measured for every k, the number of most likely words the model is allowed to choose from. The configuration term is the larger of the two at every k (Figure 2): 5.4 times larger at the $k = 40$ used by the original pipeline, and the ratio grows as the restriction loosens, from 3.7 at $k = 5$ to 7.4 at $k = 640$. Therefore, loosening the restriction does not turn the test into an identity measurement. The identity term moves too, from 0.0017 nats at raw to 0.100 at $k = 40$, so even a detector that could read it alone would still reject *Model A*'s own restricted text, albeit by a smaller margin. Almost nobody switches the restriction off: Hugging Face Transformers defaults to $k = 50$ (Hugging Face, 2026), Gemma 3 to 64 (Gemma Team, 2026) and llama.cpp to 40 (llama.cpp contributors, 2026), and the original validation did not switch it off either (Appendix B). Because the decomposition is exact by derivation, this comparison verifies our measurement procedure. The two measured terms sum to the directly computed score in all 54 generator-by-domain cells, differing by at most 1.49 standard errors (Appendix E), which confirms that the two quantities are estimated correctly.

[1] Equation (6) is an identity rather than a bound, and it requires *pure* truncation and a generator that is the evaluator; Appendix A gives both conditions and the general form.

4.2 Failure Mode 2: Insufficient Sensitivity to Model Differences

Section 4.1 varied the decoding and kept the generator fixed. This section does the reverse: every text is raw-sampled from the full vocabulary, and only the model that wrote it changes. Recall from equation (5) that expected liketropy consists of two terms. The configuration term D measures how far the generator's sampling distribution sits from the evaluator's in entropy. The identity term D_{KL} measures how far the generator's model sits from the evaluator's. The test attributes a text to *Model A* when its score is near zero. That rule is reliable only if text from every other generator scores away from zero. In this section, we conduct a set of experiments that test this mechanism.

4.2.1 LLMs as Writers (Where Ground Truth Is Observable)

We first compare *Model A* with models from its own family that share its tokenizer. The purpose of this design is to ensure that, unlike human text or text from LLMs with different tokenizers, both terms can be observed directly, position by position. We use two anchors, Llama-3-8B and Mistral-7B-v0.1, each playing the role of *Model A*. Their relatives are the next base versions (Llama-3.1-8B and Mistral-7B-v0.3) and their instruction-tuned variants. The anchor scores each relative's text with liketropy, and the AUROC measures how well that score separates the relative's text from the anchor's own raw-sampled output. At every position of the relative's text we also compute the relative's entropy $H(p^V)$, the configuration term D and the identity term D_{KL}. The difference $D - D_{\mathrm{KL}}$ reproduces each liketropy score to within 0.016 nats, so the results show what each score is made of.

Table 4: **Liketropy mainly responds to a change of entropy rather than to a change of model.** Each row is a relative of the anchor. The anchor scores the relative's raw-sampled text, and the AUROC is how well that score separates it from the anchor's own raw-sampled output. All terms are measured at the same token positions. $H(p^V)$ is the relative's entropy, D the configuration term of equation (5), so the anchor's entropy is $D + H(p^V)$, and D_{KL} is the relative-to-anchor divergence. The ratio column compares D_{KL} with the floor defined below the table.

Generator	Step from anchor	liketropy	$H(p^V)$	D	D_{KL}	D_{KL}/floor	AUROC
Anchor: Llama-3-8B	own output -0.033, $h_N(A, A) = 2.659$						
Llama-3.1-8B	version, base	-0.082	2.718	-0.014	0.055	$3.7\times$	0.582
Llama-3-8B-Instruct	post-training	$+0.542$	1.161	$+0.919$	0.361	$24.1\times$	0.984
Anchor: Mistral-7B-v0.1	own output -0.005, $h_N(A, A) = 2.204$						
Mistral-7B-v0.3	version, base	-0.034	2.196	$+0.006$	0.036	$2.4\times$	0.559
Mistral-7B-Instruct-v0.1	post-training	$+0.179$	1.661	$+0.464$	0.289	$19.3\times$	0.770
Mistral-7B-Instruct-v0.2	post-training	$+0.256$	1.499	$+0.418$	0.165	$11.0\times$	0.924
Mistral-7B-Instruct-v0.3	post-training	$+0.150$	1.779	$+0.232$	0.083	$5.5\times$	0.799

***Floor**: the smallest per-token divergence at which a text of this length can support any test.* $\log(1/\alpha)/N = 0.015$ *nats at* $N = 200$ *and the framework's own* $\alpha = 0.05$ *(Appendix G).*

Table 4 gives two results. First, a version step changes the model but not the verdict. From Llama-3 to Llama-3.1 and from Mistral v0.1 to v0.3, the entropy barely changes ($D = -0.014$ and $+0.006$), so the identity term, 0.055 and 0.036, is the whole difference between writer and anchor. Correspondingly, the AUROCs are 0.582 and 0.559, barely above chance. Second, the four instruction-tuned relatives are detected, at AUROCs from 0.770 to 0.984, but for the wrong reason. Instruction tuning narrows what a model writes, as truncated decoding does in Section 4.1, and pushes the configuration term up by 0.232 to 0.919 nats; the identity term, 0.083 to 0.361 nats, only pulls the score back towards zero.

On those four rows the test looks right and is reading the wrong thing. This is the half of the failure a better statistic could repair, since the difference between the writers is there and only the reading of it is wrong. Section 4.2.2 turns to human text.

4.2.2 HUMANS AS WRITERS

Human text is where the test fails most clearly. Averaged over the six evaluators, human text scores -0.067 and the evaluator's own text scores -0.026 (Figure 1). The two are close on all eight datasets, and neither is always the higher one (Table 19). Failing to find a difference does not show that two things are the same, so we use an equivalence test. For three of the six evaluators, it confirms that the model's own text and human text score the same (Table 9). For the other three, the score picks the right text in only 57 to 68 out of 100 pairs (AUROC 0.57 to 0.68). Longer text does not help. From 200 to 800 tokens, both sets of scores get less spread out, but their distances from zero stay within 0.018 nats of each other (Appendix M).

A score near zero does not show that human writing is close to *Model A*. The two terms can cancel, and if temperature 1 is already the model's best fit to human text, which training pushes it towards, the score is zero however far apart the two writers are (Appendix D). The failure belongs to liketropy, not to the texts: on the same texts, a supervised detector reaches AUROC 0.84 (Section 3).

The assumption meant to prevent this. The framework assumes that every other source scores away from zero. For human text this cannot be checked at deployment, since checking it needs the very distribution the test is trying to identify.

5 DISCUSSION AND LIMITATIONS

Both failures happen at one threshold. We set the threshold at 0.339 nats to accept 95% of each evaluator's own raw-sampled text (Appendix I), it rejects 78% of the evaluator's truncated text and accepts 88% of human text. This cut-off is just the noise in the scores, about 1.96 standard deviations. The framework's own theory gives no usable cut-off either. Its bounds require the two kinds of text to score 23.40 nats apart, but no score we measure reaches even 1 nat (Appendix I).

No simple repair. To use liketropy for attribution, a detector must know the configuration term. It cannot: that term depends on the writer's entropy, and the identity term depends on the writer's distribution, exactly what the test tries to find. Three ways out remain; none works well. The writer could report its decoding settings, but a writer who wants to avoid detection will not. A new score could cancel the configuration term, but no detector in Section 3 manages this, and even a perfect cancellation would fall short, since the identity term also moves with decoding (Section 4.1). Or the test could ask only whether a human or a model wrote the text, but on raw-sampled text liketropy answers this at chance (Section 4.2). Loosening the truncation does not help either (Section 4.1).

What a working guarantee needs. Every guarantee in this literature depends on how the writer chooses its words, so the party giving the guarantee must be able to see that. A detector in the log-likelihood family cannot. What its score means depends on the writer's distribution p^B, and the detector only ever sees the finished text (Section 3). Watermarking also depends on the writer, through its entropy, but that entropy is checked by the one party that can measure it (Kirchenbauer et al., 2023). The company running the model sees the entropy while the text is being written, and can decline to vouch for text when it is too low. A reliable guarantee must therefore come from the side that generates the text, not the side that reads it.

Limitations. Four limits bound the scope. Texts in XSum and RAID run to roughly 250 words, capping the convergence experiment at 800 scored tokens; a single 24 GB GPU capped model size near 8B parameters, whereas Radvand et al. (2026) report the test remaining reliable at 32B; all experiments use English text; and all six evaluators are base models, so whether RLHF-aligned models behave differently *as evaluators* deserves separate evaluation.

Mary had a little what? Ask any English speaker to finish "Mary had a little ____" and nearly everyone says "lamb". Ask a base LLM and it says the same, because it learned from us. On questions like this, the test cannot tell one writer from another. Now give one writer a short list of allowed words before it answers. Its writing style changes while the writer's identity stays the same, and the test reacts to the style instead. What the test detects is not who is writing, but how their word list was restricted.

REPRODUCIBILITY STATEMENT

Every number, table and figure in this paper is produced by one analysis notebook from stored per-document score files, and that notebook carries a regression block locking the pooled conditions, the six per-evaluator values, the 2×2 design, the four misordering counts and both operating points, so a rerun that moves any of them fails loudly rather than silently reporting a changed value. Appendix A gives the generation and scoring protocol, including the length control, the token exclusion window and the numerical precision at which every score was computed; Appendix K gives the model selection and Appendix B every constant used to instantiate the framework's bounds. The generation scripts, the per-document score files and the analysis notebook will be released.

USE OF GENERATIVE AI

In this work, we used generative AI tools to turn the authors' ideas into formal mathematical statements, including the decomposition of Proposition 1, its proof and the identity of equation (10); to help adapt the released code of Radvand et al. (2026) for generation and scoring; to implement, from the authors' specifications, the analysis notebook that produces the tables and figures; to give feedback on the experimental design and on the interpretation of results; and to propose the temperature explanation of Appendix D. We also used them to edit the writing. The research questions, the experimental design and the design of every table and figure are the authors' own. We did not use generative AI tools to generate data, clean or reformat datasets, or translate text. We checked every AI-assisted derivation by hand, tested the analysis notebook through a regression block that locks its reported values, and reviewed all AI-drafted and edited text. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. Real or fake? Learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351*, 2019.

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 *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*, 2024.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. Position: On the possibilities of AI-generated text detection. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3): 755–770, 1959.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition, 1988.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2006.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, 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 (Volume 1: Long Papers)*, pp. 12463–12492. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.674.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1082.

Alessandro Gambetti and Qiwei Han. Combat AI with AI: Counteract machine-generated fake restaurant reviews on social media. *arXiv preprint arXiv:2302.07731*, 2023.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 111–116, 2019.

Gemma Team. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Gemma Team. Gemma-3-27B-IT [model card]. HuggingFace, `https://huggingface.co/google/gemma-3-27b-it`, 2026. `generation_config.json` sets `top_k=64`; accessed 10 March 2026.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, S. Ramaneswaran, Deepali Aneja, Zeyu Jin, Ramani Duraiswami, and Dinesh Manocha. A closer look at the limitations of instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

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 *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*, 2020. URL `https://openreview.net/forum?id=rygGQyrFvH`.

Hugging Face. Generation. Transformers documentation, `https://huggingface.co/docs/transformers/en/main_classes/text_generation`, 2026. `GenerationConfig` sets `top_k=50`; accessed 10 March 2026.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1808–1822, 2020.

Mojan Javaheripi and Sébastien Bubeck. Phi-2: The surprising power of small language models. Microsoft Research Blog, `https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/`, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 219–230, 2008.

Tom Kempton, Stuart Burrell, and Connor J. Cheverall. TempTest: Local normalization distortion and the detection of machine-generated text. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1972–1980. PMLR, 2025. URL https://proceedings.mlr.press/v258/kempton25a.html.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pp. 17061–17084. PMLR, 2023.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*, 2024.

Daniël Lakens. Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017.

Guokai Li, Jiaxin Liang, Mo Liu, Yanzhe Lei, Stefanus Jasin, Fenghua Yang, and Preet Baxi. Asymptotically optimal sequential testing with heterogeneous LLMs. *arXiv preprint arXiv:2604.01086*, 2026.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. GPT detectors are biased against non-native English writers. *Patterns*, 4(7):100779, 2023.

Mingfeng Lin, Henry C. Lucas, and Galit Shmueli. Too big to fail: Large samples and the p-value problem. *Information Systems Research*, 24(4):906–917, 2013.

llama.cpp contributors. common/common.h. https://github.com/ggml-org/llama.cpp/blob/master/common/common.h, 2026. Sampling defaults set top_k=40; accessed 10 March 2026.

Jayawant N. Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 2010.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24950–24962. PMLR, 2023.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, 2018.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 309–319, 2011.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pp. 27730–27744, 2022.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*, 2022.

Tara Radvand, Mojtaba Abdolmaleki, Mohamed Mostagir, and Ambuj Tewari. A training-free method for LLM text attribution. *arXiv preprint arXiv:2501.02406v5*, 2026. Version 5, 22 March 2026.

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. URL `https://openreview.net/forum?id=OOgsAZdFOt`.

Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J. Jansen. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 64:102771, 2022.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

Lav R. Varshney, Nitish Shirish Keskar, and Richard Socher. Limits of detecting text generated by large-scale language models. *arXiv preprint arXiv:2002.03438*, 2020.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. DNA-GPT: Divergent N-Gram analysis for training-free detection of GPT-generated text. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*, 2024.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Experimental Protocol

Overview. Section 2 gives the roster, the datasets, the seven generator conditions of Table 1 and the cell counts. All six evaluators are base models; instruction-tuned variants enter as generators in the within-family condition of Appendix F, where post-training is the object of study rather than a confound. Excluding the first 30 tokens moves liketropy by at most 0.032 nats against the original treatment, as measured in the Text treatment paragraph below. Condition A_k is used in Section 4.1 and V in Appendix F.

Models. Appendix K gives the selection rationale. Mistral-7B-v0.1 doubles as the anchor of the within-family condition of Appendix F.

Datasets. XSum (Narayan et al., 2018) is carried over from the original paper for comparability. The seven RAID domains (Dugan et al., 2024) are arXiv abstracts, book summaries, news, poetry, recipes, Reddit posts and product reviews. RAID's Wikipedia domain is excluded, since Wikipedia is heavily represented in the pretraining corpora of most models studied here (Zhang et al., 2022; Penedo et al., 2023; Grattafiori et al., 2024) and memorisation would confound the distributional properties under investigation.

Generation length. A generation that falls short of its target word count is resampled, a generation-side constraint the scoring pass does not mirror, so the sampled distribution is not exactly p^A and equation (5) does not require the expected statistic to vanish. It is not, however, what displaces condition A from zero. Table 5 reports a control run with no minimum length and no resampling, otherwise identical: on the evaluator with the largest displacement the shift is $+0.001$, and the unconstrained value is still 6.9 standard errors below zero. Across the six evaluators, condition A sits between -0.005 and -0.069 rather than at zero, compared with the $+0.475$ that truncation produces. The instrumented sweep of Appendix E, where the entropy and the log-likelihood come from the single forward pass that drew each token, returns -0.014 to $+0.002$ at raw sampling, so the residual is associated with the generate-then-rescore route rather than with the statistic; those two runs differ in more than that route, so we do not attribute it further. It is an order of magnitude below every effect this paper measures, and every comparison here is between conditions scored by the same pipeline.

Table 5: Condition A against an unconstrained control. The control removes the minimum length and the resampling and changes nothing else: prompts are condition A's own excluded prefixes, the model is loaded as it was there, the end-of-sequence token is neither suppressed nor discarded, and the generated token IDs are scored directly, with no detokenisation round trip that could rescore a different segmentation. Generation is capped at 300 new tokens, which stops a sequence but never rejects and redraws one, so it constrains the length distribution without biasing the per-token statistic. XSum, $n = 400$ per cell. Intervals are standard errors; the shift's interval is the standard error of the difference between two independent runs.

Evaluator	**Condition** A	**Uncon-strained**	**Shift**	**Scored tokens** median	mean	**At cap**
Falcon-RW-1B	-0.0216 ± 0.0092	-0.0080 ± 0.0105	$+0.0136 \pm 0.0139$	299	250	283/400
Gemma-7B	-0.0530 ± 0.0079	-0.0521 ± 0.0076	$+0.0009 \pm 0.0109$	299	264	298/400

Text treatment. The original paper computes liketropy over all tokens from the start of the generation, even though the first 30 tokens are the human-written prompt, and compensates by switching to a one-sided test (Radvand et al., 2026, §7.2), which assumes the bias shifts both conditions equally and sacrifices two-sided power rather than removing the contamination. We preserve that treatment for both double-primed conditions, A'' and B'', which reproduce the original setup. Excluding the prompt elsewhere leaves a sum over tokens that are entirely model-generated when the generator is a model and entirely human-written when it is not. The sweep A_k reaches the same window by construction rather than by skipping: its instrumented run forces generation to exactly 200 new tokens and scores those alone, which is why its raw-sampling row returns -0.014 to $+0.002$, whereas the companion sweep, whose window keeps the prompt, returns -0.018 to -0.070 on the same models. Length is controlled by word count for fairness across tokenizers, while the prompt and exclusion window are measured in tokens because they operate where the model processes input. We adopt the stricter treatment wherever human text is involved, since there the shared prefix would otherwise be common to both arms. The size of the effect is measurable on XSum, where a companion full-vocabulary run scored under the original treatment differs from condition A in the scoring window alone: across the six evaluators the two differ by 0.005 to 0.032 nats, about one fifteenth of the A-to-A'' shift of $+0.468$ on the same dataset.

When equation (6) holds. The closed form $D_{\mathrm{KL}}(\widetilde{p}_n^A \| p_n^A) = -\log Z$ is an identity, not a bound, and it needs two things. It needs *pure* truncation: a decoding step that reshapes the retained set destroys the constant ratio on which it depends, and at temperature 0.7 on top of top-k the divergence is roughly twice $-\log Z$, which is why the sweep pins temperature at 1.0 with no repetition penalty or n-gram blocking. It also needs the truncated generator to be the evaluator. For a foreign generator C truncated to $\widetilde{p}^C$ and scored by A,

$$D_{\mathrm{KL}}(\widetilde{p}_n^C \,\|\, p_n^A) \;=\; -\log Z \;+\; \underbrace{\frac{1}{Z}\sum_{y \in S} p_n^C(y) \log \frac{p_n^C(y)}{p_n^A(y)}}_{D_S,\ \text{the } C \to A \text{ divergence over the retained set}} \quad , \tag{7}$$

which recovers condition B when S is the whole vocabulary, since then $Z = 1$ and D_S is the full divergence, and recovers equation (6) when $C = A$, since then $D_S = 0$. Equation (6) is therefore the $C = A$ instance of equation (7) rather than a coincidence. It also explains why the split of Section 4.1 is available only for self-generation: truncating a foreign generator moves the identity term by $-\log Z$ plus an unobservable D_S, and moves the configuration term by the entropy removed from C rather than from A. The interaction term of the 2×2 design, -0.054 ± 0.009, is that difference. Finally, the reversed direction $D_{\mathrm{KL}}(p_n^A \| \widetilde{p}_n^A)$ is infinite, since p_n^A places mass where $\widetilde{p}_n^A$ places none; the framework's conditions are written generator-first, which is the finite direction.

The instrumented sweep. Resolving both terms of equation (5) requires the pre-truncation and post-truncation distributions at every position, so the sweep comes from a dedicated instrumented run on XSum in which generation was forced to exactly 200 new tokens, suppressing the end-of-sequence token, and documents were filtered by token count rather than word count. Appendix E returns to the consequences.

Numerical precision. Weights are loaded with 8-bit quantisation, which is what lets all six evaluators fit on a single 24 GB GPU. Scores are computed in float32 throughout, with one exception: in the length sweep of Appendix M, Falcon-RW-1B and OPT-1.3B are scored in bfloat16, so their per-document values fall on a 0.0156 grid. The rounding this adds has standard deviation 0.0045 against standard deviations above 0.08 in those rows, and is below the precision reported here. Generation and scoring use the same quantised model in every condition, so equation (5) splits the quantised distribution rather than straddling two, and the within-family verdict of Appendix F does not turn on how closely quantised entropies track unquantised ones: the two version steps are separated at AUROC 0.582 and 0.559, and no quantisation effect of the size measured here moves a value of that order.

B Constants of Radvand et al.'s Theorem 1

All quotations and all section, table, footnote and theorem numbers of Radvand et al. (2026) in this paper refer to version 5 (arXiv:2501.02406v5, 22 March 2026); earlier versions appeared under a different title.

Decoding in the original validation. Radvand et al.'s reported figures come from truncated generation, so their foreign-generator condition is our B'' at $+0.270$ rather than our B at -0.285: the sign flips while the magnitude stays close, so the separation reported there comes mainly from the configuration term.[2] This is the same mechanism that places truncated text outside the premise of Theorem 1(a), so the two validations cannot be repaired independently.

The null hypothesis and their Example 1. The null hypothesis is broader than the premise of their Theorem 1(a), since Theorem 1(a) proves concentration for text *sampled from* p^A while the claim that *Model A* generated the text covers decoding as well. Radvand et al. (2026) take the opposite view by example, calling a greedy variant of *Model A* "the unknown generator B": a reading that concedes the test attributes text to a model paired with a decoding configuration rather than to a model. Their Example 1 is the same effect seen from the other side, with greedy *A*'s text drawing lower log-perplexity under *A* than *A*'s own sampled output, and the entropy anchor is introduced to repair it. Equation (5) says why the repair is partial: concentration enters through D, which the anchor does not bound, so under truncation it carries the evaluator's own text further from zero than foreign text rather than nearer to it.[3] Theorem 1 fixes no magnitudes: its scale constants c_1 and c_2 are only asserted to exist, so the bounds carry no numerical content until they are instantiated, and Radvand et al. (2026) report no numerical threshold anywhere in the paper. Table 6 lists the constants, instantiated for Gemma-7B, the largest-vocabulary evaluator in the study and so the one

[2] `Whiteboxdetection.ipynb`, the code their Appendix F.9 gives for §7.2, sets `top_k` $= 40$ and `top_p` $= 0.96$, and its nine experiment cells reproduce their Table 5 cell by cell, so these are the settings the code requests. Its sampling routine applies one flag or the other, never both, so generation runs at top-$p = 0.96$ with Hugging Face's default `top_k` $= 50$; no branch reaches the full vocabulary.

[3] They report the same direction empirically, greedy generations being easier to detect than sampled ones (Radvand et al., 2026, §7.3).

that makes the bounds hardest to meet. Its tokenizer defines $256{,}000$ tokens, the figure Table 17 reports, while its output layer is padded to $256{,}128$; the softmax is normalised over the padded width, so that is the K used here and in Appendix G. The two differ by 0.0005 nats in $\log K$ and change nothing downstream. Our instantiation is $c_1 = 2$ and $c_2 = 4$, identified with the sub-exponential norms (Vershynin, 2018) computed in their Lemmas 3 and 5 ($2 \log K$ and $-4 \log \varepsilon$), the tightest reading the proofs support up to absolute constants absorbed by the concentration inequality. For the minimum token probability ε of Assumption 1 we adopt the value Radvand et al. (2026) supply, 10^{-16}, which they justify by reference to a default optimiser constant. The choice is theirs, and it is unfavourable to their own bounds: the quantity entering Theorem 1(b) is $-c_2 \log \varepsilon$, so moving from 10^{-6} to 10^{-16} enlarges every threshold by a factor of $8/3$, exactly and in any log base. Two cautions attach to their justification, neither of which changes our numbers. The constant they cite governs numerical stability in the denominator of an adaptive optimiser update rather than a floor on next-token probabilities, and they report it as $-\log_{10}(\varepsilon) = 16$ while the bounds are written in natural logarithms, where the value is 36.8. Assumption 1 is imposed on the generator as well as the evaluator, and B may be a human writer, for whom a floating-point argument about model software gives no reason a next-token probability cannot fall below ε.

Table 6: Constants of their Theorem 1 and of the conditions it relies on, with the values used in this paper.

Symbol	Quantity	Enters through	Value used here (provenance)
K	output-layer width	$\nu = c_1 \log K$	256,128 (Gemma-7B, observable)
ε	minimum token probability (Assumption 1)	$\nu = -c_2 \log \varepsilon$	10^{-16} (theirs, Remark 2)
c_1	scale constant of Theorem 1(a)	$\nu = c_1 \log K$	2 (Lemma 3: norm $2 \log K$)
c_2	scale constant of Theorem 1(b)	$\nu = -c_2 \log \varepsilon$	4 (Lemma 5: norm $-4 \log \varepsilon$)
ε_1	divergence slack, model–model (Assumption 2)	their Prop. 1	slack; no instantiation needed
ε_2	separation slack (Assumption 3)	their Prop. 2	slack; no instantiation needed
c_3	auxiliary constant (Lemma 6)	$0 < c_3 < \varepsilon_2$	not instantiated

C Proof of Proposition 1 and Related Conditions

Proof of Proposition 1. Since α_n, β_n and q_n are fixed given $y_{<n}$, $\mathbb{E}[\alpha_n \log q_n(y_n) + \beta_n \mid y_{<n}] = \beta_n - \alpha_n H(p_n^B, q_n)$, and the cross-entropy identity $H(p_n^B, q_n) = H(p_n^B) + D_{\mathrm{KL}}(p_n^B \| q_n)$ (Cover & Thomas, 2006, Thm. 5.4.3) gives the claim termwise. Taking the outer expectation over prefixes drawn from B leaves the form unchanged.

The same split, applied to the wider literature. The connection between a divergence and detection error is not itself new: Varshney et al. (2020) give the Neyman–Pearson error exponent for distinguishing human text from a model's as $\exp(-N D_{\mathrm{KL}}(p^H \| p^A))$. What equation (4) adds is that the quantity a detector in this family observes is not that divergence but its combination with an entropy, and that the two cannot be read apart from one score. Separation conditions stated elsewhere in the literature can be read the same way. Yang et al. (2024) require the expected log-likelihood of machine text under a fixed reference to exceed that of human text by a margin $\Delta > 0$, and derive from it the number of re-promptings their method needs. For any reference q admitted by Definition 1, Proposition 1 expands that margin as

$$\mathbb{E}_{p^M}[\log q] - \mathbb{E}_{p^H}[\log q] = \underbrace{\big[H(p^H) - H(p^M)\big]}_{\text{configuration}} + \underbrace{\big[D_{\mathrm{KL}}(p^H \| q) - D_{\mathrm{KL}}(p^M \| q)\big]}_{\text{identity}}, \tag{8}$$

whose first bracket does not depend on which model q is.[4] What such a guarantee requires is therefore not a margin in identity, and part of what sets it is the generator's decoding. Appendix J measures the consequences for the detectors of Table 15.

[4] Yang et al. (2024) do not fix q where they state that inequality; equation (8) holds for every admissible choice, so the reading does not turn on the identification.

D WHY HUMAN TEXT SCORES NEAR ZERO

Section 4.2.2 reports that human text scores about where *Model A*'s own text does. A score near zero does not imply a small divergence. By equation (5), a zero mean says only

$$\mathbb{E}\big[H(p^A) - H(p^H)\big] \;=\; \mathbb{E}\big[D_{\mathrm{KL}}(p^H \,\|\, p^A)\big], \tag{9}$$

so the two terms can both be large and cancel. Two routes lead to a score near zero, and we do not test which one operates.

Route 1: training. Maximum-likelihood training minimises the cross-entropy $H(p^H, p^A) = H(p^H) + D_{\mathrm{KL}}(p^H \| p^A)$, so it lowers the identity term, and if that term reached zero the configuration term would follow. This route is enough but not needed, and nothing measured here shows that the divergence is small.

Route 2: temperature. Let $z^A(x)$ be the evaluator's logits at a human prefix x, let $p^A_s(\cdot \mid x) = \mathrm{softmax}(s\, z^A(x))$, and let $L_H(s) = \mathbb{E}_{(x,y)\sim p^H}[\log p^A_s(y \mid x)]$ be the expected log-likelihood of human text at temperature $1/s$. Differentiating at $s = 1$,

$$L'_H(1) \;=\; \mathbb{E}_{p^H}\big[z^A_y(x) - \mathbb{E}_{Y\sim p^A(\cdot|x)} z^A_Y(x)\big] \;=\; \mathbb{E}_{p^H}\big[\log p^A(y \mid x) + H(p^A(\cdot \mid x))\big] \;=\; \mathbb{E}_{p^H}[\text{liketropy}]. \tag{10}$$

If $s = 1$ is already the best temperature for human text, expected human liketropy is therefore exactly zero, whatever the divergence. At a minimum of the training loss, rescaling all logits cannot lower that loss, so $s = 1$ is stationary on the training distribution; on a new domain this holds only approximately. The sign also has a reading. The human mean measured here is negative, so $L'_H(1) < 0$: the evaluators are slightly overconfident on these domains and would fit human text better at a temperature above one.

E THE DECOMPOSITION AT $k = 40$, PER MODEL

Table 7: The two terms per model at $k = 40$, with the falsification check. $D = H(p^A) - H(\widetilde{p}^A)$ and $D_{\mathrm{KL}} = -\log Z$, the closed form of equation (6), which agrees with the direct sum over the vocabulary to within 5.2×10^{-7} across all 54 cells. *Predicted* is $D - D_{\mathrm{KL}}$, computed from the two distributions; *observed* is the mean liketropy of the sampled tokens. The two are obtained independently, so their agreement tests the decomposition; Δ reports |predicted − observed| in units of the standard error of the observed mean. The Mean row's ratio is the mean of the six per-model ratios and its interval their standard error; bootstrapping documents within a cell puts the sampling error on each per-model ratio at about 0.04, so the spread across models is dispersion between models rather than noise. XSum, $n = 200$ texts per cell.

Evaluator	Inputs			The two terms			Verification		
	$H(p^A)$	$H(\widetilde{p}^A)$	Z	D	D_{KL}	D/D_{KL}	predicted	observed	Δ (SE)
Falcon-RW-1B	2.615	1.907	0.875	+0.709	+0.133	5.31	+0.575	+0.579	0.29
OPT-1.3B	2.441	1.797	0.889	+0.644	+0.117	5.49	+0.527	+0.530	0.23
Phi-2	2.388	1.795	0.895	+0.593	+0.111	5.34	+0.482	+0.493	0.81
Gemma-7B	2.195	1.723	0.921	+0.472	+0.082	5.74	+0.390	+0.393	0.24
Llama-3-8B	2.232	1.770	0.915	+0.462	+0.089	5.20	+0.373	+0.384	0.89
Mistral-7B-v0.1	1.958	1.565	0.932	+0.393	+0.070	5.62	+0.323	+0.308	1.35
Mean	2.305	1.759	0.905	**+0.545**	**+0.100**	**5.45 ± 0.08**	+0.445	+0.448	0.63

The sweep neutralises top-p, so these are top-k-only values and the retained mass is higher than under the original paper's combined setting: the two truncation filters are applied in sequence, so top-$p = 0.96$ discards the least probable 4% of the mass top-k leaves rather than 4% of the full distribution. The run also generates to a fixed length on XSum alone, so its absolute level is not directly comparable with the eight-domain results of Figure 1; it is the ratio between the terms that carries forward. At raw sampling both terms are zero in expectation, and what is measured is

sampling noise at the 10^{-3} nat scale: the configuration term comes out at -0.004 for Llama-3-8B, below the zero it must take, and the per-model ratio scatters from -0.98 to $+1.20$ against 5.20 to 5.74 at $k = 40$. That is why the ratio is not reported for the raw row. Two figures the main text quotes are read off the whole sweep rather than off this table. The largest disagreement between predicted and observed anywhere in the 54 cells is Phi-2 at $k = 80$, at 1.49 standard errors, larger than any entry above. Finally, D/D_{KL} roughly doubles within every model, from 3.58–3.77 at $k = 5$ to 6.22–7.81 at $k = 640$ (a factor of 1.74–2.14), against 5.20–5.74 across the six models at $k = 40$.

F The Within-Family Experiment

What the test detects within a family is configuration. Table 4 in Section 4.2.1 splits each generator's liketropy into the two terms. The mechanism is entropy collapse: preference optimisation (Ouyang et al., 2022) narrows the output distribution, an effect reported independently of detection (Kirk et al., 2024; Ghosh et al., 2024). Measured at the same token positions, post-training lowers entropy from the anchor's 2.080 to 1.161 nats for Llama-3-8B-Instruct and from 2.125 to 1.661 for Mistral-7B-Instruct-v0.1 ($D + H(p^V)$ against $H(p^V)$ in Table 4), the same narrowing that top-k produces in Section 4.1. The two base-to-base version steps leave it where it was, their configuration terms being -0.014 and $+0.006$ nats, and their liketropy averages -0.058 nats (Table 11). A larger base-to-base step bounds the same axis from the other end. Llama-2-7B, a full model generation behind its anchor rather than one version, scores -0.302, close to the -0.285 of the unrelated generators in condition B: one generation back, a relative is no longer distinguishable from a stranger by this statistic, so the identity signal is not merely weak but saturates well inside the family. It is held out of the pooled row and of Table 4 because its SentencePiece $32{,}000$ vocabulary is not commensurable with the anchor's tiktoken $128{,}256$ vocabulary, which the divergence split requires, though the liketropy itself needs no shared tokenization.

G The Feasibility Floor

This appendix asks whether the divergences in Table 4 are large enough for any test to use. With $I(B\|A) := \frac{1}{N}\sum_n D_{\mathrm{KL}}(p_n^B \| p_n^A)$ denoting the per-token information rate, the standard rate-to-sample-complexity relation (Wald, 1945; Chernoff, 1959; Li et al., 2026) puts the tokens needed to reach error probability α at order $\log(1/\alpha)/I(B\|A)$, so a test confined to N tokens is feasible only if

$$I(B \,\|\, A) \;\gtrsim\; \frac{\log(1/\alpha)}{N}. \tag{11}$$

Radvand et al. (2026) reach the same floor independently. Their Corollary 1 gives $\alpha_N(T) + \beta_N(T) \geq \frac{1}{2} e^{-N D_{\mathrm{KL}}(P_B \| P_A)}$ for any test T, which, with both error probabilities held at α, rearranges to $D_{\mathrm{KL}} \geq \log(1/(4\alpha))/N$, the same order as equation (11) and within a factor of two of it at the settings tabulated below. The criterion applied here is therefore one the framework itself endorses. Two cautions attach. It is an order-of-magnitude relation derived for independent observations and known rates, so margins of a small factor either way carry little weight; and it constrains the mean per-token rate only, whereas detectability depends on that rate relative to the standard deviation of the per-token increments. The floor therefore certifies impossibility below it without certifying feasibility above it. It is insensitive to α, which enters logarithmically: tightening α from 0.05 to 10^{-6} raises the floor by less than a factor of five at any fixed length. At the framework's own 95% convention and the longest scored length in this study, feasibility requires $I \gtrsim 3.7 \times 10^{-3}$ nats per token (Table 8). Every divergence in Table 4 clears the floor at the length it was measured at, $N = 200$, where the floor is 0.015 nats per token: by 2.4 and 3.7 times on the two version steps and by 5.5 to 24.1 times on the four post-training steps, which is the second-to-last column of that table. The information needed to separate those generators is present; what fails is the statistic's ability to use it. The reading is model-against-model only. Human text is the case where the divergence is not observed (Appendix D), and nothing here says whether it clears the floor.

Table 8: Feasibility floor $\log(1/\alpha)/N$ (nats per token) below which the information rate cannot support detection within N tokens at error tolerance α.

N	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 10^{-3}$	$\alpha = 10^{-6}$
200	0.0150	0.0230	0.0345	0.0691
300	0.0100	0.0154	0.0230	0.0461
400	0.0075	0.0115	0.0173	0.0345
600	0.0050	0.0077	0.0115	0.0230
800	0.0037	0.0058	0.0086	0.0173

H Statistical Tests

Misorderings, cell by cell. Under the two-sided rule a misordering occurs whenever the wrong generator sits nearer zero. Own truncated text is further from zero than foreign truncated text in all 240 cells, than human text in all 48, and than foreign raw text in 209 of 240 (Table 10). The one comparison the rule gets right is $|A| < |B|$, which holds in all 240 cells, and it is the regime where both generators sample from full distributions. On the four pooled means of Figure 1 the sampling manipulation moves the statistic 2.3 times as far as the identity manipulation (bootstrap 95% CI 2.1 to 2.5 over the 48 cells): $\frac{1}{2}[(A''\text{–}A)+(B''\text{–}B)] = +0.528$ against $\frac{1}{2}[(B\text{–}A)+(B''\text{–}A'')] = -0.232$, though neither contrast isolates a term of equation (5).

Equivalence of condition A and human text. Showing that two distributions coincide requires an equivalence test rather than a difference test, since at $n = 3{,}200$ a t-test rejects equal means for negligible differences (Lin et al., 2013) and failing to reject is not evidence of sameness. Using Cohen's d (Cohen, 1988), parametric AUROC (Hanley & McNeil, 1982) and TOST (Schuirmann, 1987; Lakens, 2017) at a margin of $0.2\times$ pooled SD, we certify equivalence to human text for the own output of three of the six base evaluators, with d between 0.033 and 0.066 and AUROC between 0.509 and 0.519. Mistral-7B-v0.1, Phi-2 and OPT-1.3B separate more (d 0.257, 0.372 and 0.651; AUROC 0.572, 0.604 and 0.677) and are not certified, though an AUROC below 0.7 is still poor classification (Mandrekar, 2010). The equivalence test is clustered on the eight domains, which are the unit over which the design is replicated: treating the 3,200 texts as independent understates its standard error by a factor of 2 to 4, in the direction that makes equivalence easier to certify. Table 9 gives the full statistics. The 31 exceptions to $|A''| > |B|$ are concentrated rather than scattered:

Table 9: Equivalence test of each evaluator's own raw-sampled output against human text, with the first 30 tokens excluded, pooled over all eight datasets. TOST equivalence margin of $0.2\times$ pooled SD; small p certifies equivalence, p near 1 means it cannot be certified. The test is clustered on the eight domains, the per-domain difference being the unit of analysis, so it is t with seven degrees of freedom rather than a pooled two-sample test over texts. Cohen's d and AUROC are effect sizes and are unaffected.

Evaluator	A	**Human**	Δ	**Cohen's d**	**AUROC**	**TOST p**
Falcon-RW-1B	-0.0235 ± 0.1744	-0.0109 ± 0.2072	-0.0126	0.0660	0.5186	0.015
OPT-1.3B	-0.0219 ± 0.1687	-0.1551 ± 0.2351	$+0.1332$	0.6510	0.6774	0.999
Phi-2	-0.0151 ± 0.1702	-0.0872 ± 0.2146	$+0.0721$	0.3720	0.6037	0.987
Gemma-7B	-0.0690 ± 0.1868	-0.0752 ± 0.1913	$+0.0062$	0.0329	0.5093	0.026
Llama-3-8B	-0.0209 ± 0.1618	-0.0288 ± 0.1882	$+0.0079$	0.0450	0.5127	0.043
Mistral-7B-v0.1	-0.0054 ± 0.1458	-0.0470 ± 0.1767	$+0.0417$	0.2570	0.5721	0.841

16 fall on recipes, 6 on news, 5 on XSum and 4 on arXiv abstracts, and none on the four highest-entropy domains. By evaluator, they fall on Gemma-7B (11), Mistral-7B-v0.1 (8), Llama-3-8B (8), Phi-2 (3) and OPT-1.3B (1), with none on Falcon-RW-1B, and they track the margin between the two quantities the comparison puts side by side: $|A''| - |B|$ is 0.105 for Gemma-7B and 0.107 for Mistral-7B-v0.1 against 0.422 for Falcon-RW-1B (Tables 18 and 20). Both breakdowns follow the mechanism of Section 4.1: the exceptions arise where truncation removes the least entropy, whether because the domain is formulaic or because the evaluator's own distribution is already narrow.

Table 10: Misordering rates under the two-sided rule, cell level. Six base evaluators against the five other base models and against human text, over eight datasets: 240 cross-model cells and 48 human cells, 288 in all. The last row is the one ordering the rule never gets wrong.

Comparison	Cells	Rate	Reading
$\lvert A''\rvert > \lvert B''\rvert$	$240/240$	100%	own truncated text further from zero than foreign truncated text
$\lvert A''\rvert > \lvert\text{Human}\rvert$	$48/48$	100%	own truncated text further from zero than human text
$\lvert A''\rvert > \lvert B\rvert$	$209/240$	87%	own truncated text further from zero than foreign raw text
$\lvert A\rvert > \lvert B\rvert$	$0/240$	0%	own raw text further from zero than foreign raw text

I Threshold and Operating-Point Analysis

The conditions numerically. Table 11 is the numeric companion to Figure 1, and adds the acceptance rate each condition draws under each of the two decision rules at the thresholds calibrated below. The one-sided rule, which Radvand et al. (2026, §7.2) use for detection, keeps the two costly errors: it accepts only 3.8% of condition A, the evaluator's own full-vocabulary output and the case their Theorem 1(a) is stated for, and it accepts a majority of both A'' (89.6%) and B'' (54.5%), although the evaluator wrote one and not the other.

Table 11: The five generator conditions pooled over the six base evaluators and eight datasets, with two within-family conditions below, ordered by distance from zero. *2-sided* and *1-sided* are acceptance rates under the two rules of Section 2, at thresholds 0.339 and 0.250, the latter being $-t$; the 95.0% and 5.0% entries are the calibration points that set them. *Std* is the pooled per-document standard deviation, not the standard error of the mean. V ver. and V post. are the base-version and post-training relatives of the two family anchors, on XSum alone, and are discussed in Appendix F.

Cond.	By A?	liketropy	Std	2-sided	1-sided
A	yes	-0.026	0.170	95.0%	3.8%
Human	no	-0.067	0.208	88.5%	5.0%
B''	no	$+0.270$	0.182	64.6%	54.5%
B	no	-0.285	0.243	60.8%	0.8%
A''	yes	$+0.475$	0.177	22.3%	89.6%
Within family, XSum only					
V ver.	no	-0.058	0.155	–	1.8%
V post.	no	$+0.282$	0.231	–	51.4%

For each evaluator, τ is set so that 95% of its own raw-sampled generations fall inside $\pm\tau$, and the same τ is then applied to every other condition. Acceptance rates are counted from the per-document scores; fitting a normal distribution to the pooled mean and standard deviation instead moves them by up to 2.7 percentage points, since none of the five conditions is normally distributed. The threshold

Table 12: Acceptance rates at a threshold τ calibrated per evaluator to accept 95% of condition A, with $1.96\,\sigma_A$ alongside for comparison. The A column is 95.0% by construction and records that the calibration held exactly. The mean is over all six base evaluators.

Evaluator	τ	$1.96\,\sigma_A$	A	A''	Human	B	B''
Falcon-RW-1B	0.346	0.342	95.0%	7.9%	91.0%	78.7%	39.4%
OPT-1.3B	0.338	0.331	95.0%	6.8%	78.2%	45.5%	65.3%
Phi-2	0.336	0.334	95.0%	24.8%	85.6%	64.8%	60.7%
Gemma-7B	0.408	0.366	95.0%	49.2%	94.8%	68.0%	81.2%
Llama-3-8B	0.328	0.317	95.0%	23.0%	91.3%	52.5%	70.2%
Mistral-7B-v0.1	0.284	0.286	95.0%	23.8%	88.8%	53.4%	70.5%
Mean	0.340	0.329	95.0%	22.6%	88.3%	60.5%	64.5%

τ tracks 1.96 times the standard deviation of condition A closely, 0.284 to 0.408 against 0.286 to

0.366, so it is the evaluator's own self-scoring noise rather than a framework quantity; Gemma-7B is the one evaluator whose empirical quantile runs appreciably wider than the normal approximation, and Mistral-7B-v0.1's lower τ follows from its narrower spread. A single pooled threshold of 0.339, the one Table 11 uses, gives rates within half a percentage point of the per-model figures above at every condition, so the result does not depend on per-model calibration. We report per-model τ because it is the setting more favourable to the framework.

Separation of the evaluator's own output from human text. Table 13 gives the measurement Section 2 reports as a range. It is rank-based and so independent of τ, which is why it is tabulated separately from the acceptance rates above rather than as two more columns of Table 12. Table 14

Table 13: Machine-versus-human AUROC on XSum for each evaluator's own output, under the original pipeline's decoding (A'') and under raw sampling (A). Rank-based, $n = 400$ per side. Pooled over the eight datasets, the two columns run 0.982 to 0.997 and 0.478 to 0.675, so the collapse is not an artefact of the dataset the original paper reports. Table 15 puts the same statistic at 0.988 and 0.517 for Gemma-7B under its own protocol, which fixes one scorer, pools the eight datasets and skips no tokens; the two are not directly comparable and neither is a check on the other.

Evaluator	A''	A
Falcon-RW-1B	0.997	0.387
OPT-1.3B	0.999	0.555
Phi-2	0.991	0.534
Gemma-7B	0.994	0.511
Llama-3-8B	0.987	0.359
Mistral-7B-v0.1	0.993	0.476
Mean	**0.994**	**0.471**

gives the thresholds the concentration bound itself prescribes, obtained by solving both bounds of Theorem 1 for t under the instantiation of Appendix B. Their size is expected rather than anomalous: Azuma–Hoeffding-type inequalities are known to be loose for sequential processes with bounded increments (Boucheron et al., 2013). Nor does the verdict turn on ε (Appendix B), since every empirical score in this study falls below 1 nat. Both bounds have to hold at once for the test to

Table 14: Concentration thresholds t (in nats) at confidence $1 - \delta$ ($N = 200$, $K = 256{,}128$, $c_1 = 2$, $c_2 = 4$, $\varepsilon = 10^{-16}$).

Confidence	Thm. 1(a)	Thm. 1(b)
90%	3.05	18.04
95%	3.38	20.01
99%	4.05	23.99
99.9%	4.86	28.73

separate the two hypotheses, so at 95% the two conditions must be $3.38 + 20.01 = 23.40$ nats apart, which is the figure Section 5 quotes against measured scores that never reach 1 nat.

J The Decomposition Against Neighbouring Detectors

The measurement. Table 15 is the numeric form of Table 3, with the self-scoring cell separated from the mean over generators.

Table 15: **Changing only the generator's decoding moves the five detectors of Table 2 and leaves the two supervised detectors.** One scorer, Gemma-7B, playing the role of the evaluator, discriminates machine from human text over six generators and eight datasets, $n = 400$ per cell, under two generator configurations: *truncated* at top-k 40, top-p 0.96, and *raw* at full vocabulary. Nothing else changes between the two halves of the table. *Self* is the Gemma-on-Gemma cell, *mean* the average over all six generators. An AUROC below 0.5 is not chance but inversion: raw-sampled machine text scores as more human-like than human text does. The first block lists the five detectors of Table 2 in the same order; the indented row is liketropy scored under the framework's own decision rule, |liketropy|, and is not a separate detector. The second block lies outside Definition 1, since rank is not affine in $\log q_n$; it is reported because GLTR (Gehrmann et al., 2019) scores tokens by rank. *log-likelihood*, *entropy*, *rank* and *log-rank* are zero-shot baselines of Mitchell et al. (2023), each reported there with an AUROC of its own, and log-likelihood and rank also appear in the released output of Radvand et al. (2026); *liketropy* is the statistic of Radvand et al. (2026), *Fast-DetectGPT* is Bao et al. (2024) and *Binoculars* Hans et al. (2024), with Falcon-7B as observer and Falcon-7B-Instruct as performer; the supervised rows are the released GPT-2 output detectors of Solaiman et al. (2019). †*Why the two liketropy rows carry no mean.* Liketropy does not ask whether a text is machine-written; it asks whether *this scorer* wrote it. Against a foreign generator both arms of a machine-versus-human comparison are texts the scorer did not produce, so its own hypothesis answers "not A" to both and an AUROC between them measures nothing it claims. Only the self column asks it the question it was built for, and that column is also what Radvand et al. (2026, §7.2) run: their machine arm is generated by the evaluator model itself. The remaining rows are machine-versus-human detectors, for which the cross-model mean is the advertised use. The protocol differs from the rest of the paper and these numbers are not comparable with Figure 1; the protocol is given below, and Table 16 splits the raw columns by generator.

Detector	**Truncated generator**		**Raw generator**	
	self	mean	self	mean
The five detectors of Table 2				
log-likelihood	0.882	0.758	0.285	0.184
entropy	0.445	0.491	0.764	0.810
liketropy†	0.988	–	0.517	–
two-sided rule†	0.110	–	0.530	–
Fast-DetectGPT	0.990	0.933	0.528	0.295
Binoculars	0.948	0.960	0.344	0.368
Rank-based, outside Definition 1				
rank	0.788	0.722	0.309	0.196
log-rank	0.909	0.789	0.283	0.179
Supervised, surface features				
RoBERTa-base detector	0.829	0.880	0.743	0.783
RoBERTa-large detector	0.810	0.869	0.842	0.883

Section 3 claims that the decomposition reaches beyond liketropy to any statistic built from the gap between observed and expected surprisal. This appendix is the measurement behind that claim. It tests the claim by holding one evaluator fixed and changing only the generator's decoding, which is the manipulation equation (5) says moves the configuration term while moving the identity term far less. The protocol differs from the rest of the paper and its numbers are not comparable with Figure 1. A single scorer, Gemma-7B, discriminates machine from human text over the six base generators and eight datasets, $n = 400$ per cell, with no token skipped, so the Gemma column is white-box self-scoring and the other five are cross-model. Two generation settings are run over the same generators: *truncated*, at top-k 40 and top-p 0.96, and *raw*, at full vocabulary. DetectGPT is omitted, since its perturbation pass needs a T5 forward pass for every candidate.

Table 3 in Section 3 summarises the result; this appendix gives the protocol behind it, the reading of each block, and the split by generator. Four things follow. First, the collapse spares none of the detectors in the first block that work under truncation: Binoculars is the strongest in the truncated tree and still falls below 0.5 in the raw one, so normalising one surprisal gap by another does not cancel the configuration term, which closes the second of the three repair directions listed in Sec-

tion 5 empirically rather than by argument. The one entry that moves the other way is raw entropy, which rises from 0.491 to 0.810: full-vocabulary sampling is what widens the observed distribution, so there the configuration term is the whole of the signal rather than a confound within it. Second, the supervised detectors barely move on the same texts, and the larger one is slightly better on raw text than on truncated text. Raw-sampled machine text is therefore detectable; the family the decomposition covers is what cannot detect it, which places this failure between the impossibility ceiling of Sadasivan et al. (2025) and the sample-complexity possibility results of Chakraborty et al. (2024). Third, the framework's own two-sided rule is the weakest entry in the table, at 0.110 on the one cell it was designed for, the scorer judging its own output, where the signed form of the same statistic reaches 0.988. Fourth, the collapse is not an artefact of the smallest models' raw output degrading (sample continuations are in Appendix N), though it does deepen as the generator gets smaller: the five detectors that work under truncation average 0.162 on the two smallest generators against 0.294 on the two largest (Table 16). What rules out degeneracy is the fluent end of the range. Llama-3-8B and Mistral-7B-v0.1 produce raw output no one would call degenerate, and there the five still fall from 0.72–0.96 under truncation to 0.22–0.41 under raw sampling, which is the collapse the pooled figures report. The supervised rows make a different point: reaching 0.956 on Falcon-RW-1B's raw output, they show that even the most damaged text here carries ample signal for a surface classifier, so what fails on it is the family the decomposition covers and not detection as such. Radvand et al.

Table 16: The raw-generator columns of Table 15, split by generator, with rows grouped as there. Mean AUROC over the eight datasets, $n = 400$ per cell. Generators are ordered by nominal parameter count. Gemma-7B is the fixed scorer, so that column is white-box self-scoring and the other five are cross-model; it is shown for completeness and excluded from the size comparison in the text, which uses the four cross-model generators at the ends of the range. The two liketropy rows carry the self column only, for the reason given in the note to Table 15.

Detector	Falcon 1B	OPT 1.3B	Phi-2 2.7B	Gemma 7B *(self)*	Mistral 7B	Llama-3 8B
The five detectors of Table 2						
log-likelihood	0.067	0.110	0.146	0.285	0.267	0.227
entropy	0.912	0.875	0.834	0.764	0.717	0.758
liketropy	–	–	–	0.517	–	–
two-sided rule	–	–	–	0.530	–	–
Fast-DetectGPT	0.140	0.221	0.215	0.528	0.346	0.320
Binoculars	0.282	0.402	0.376	0.344	0.411	0.394
Rank-based, outside Definition 1						
rank	0.098	0.135	0.149	0.309	0.266	0.219
log-rank	0.063	0.104	0.141	0.283	0.262	0.221
Supervised, surface features						
RoBERTa-base detector	0.873	0.837	0.848	0.743	0.680	0.717
RoBERTa-large detector	0.956	0.950	0.927	0.842	0.793	0.827

(2026, §7.3) report the same direction in a black-box robustness study on RAID, greedy generations being easier to detect than sampled ones because greedy decoding concentrates probability mass on high-likelihood tokens, and conclude that combining log-perplexity with an entropy-based statistic is less brittle to decoding than single-cue detectors. That study runs on a different axis from this one, greedy against sampled rather than truncated against raw, at TPR at 5% FPR rather than AUROC, and on generations they did not produce, so the two do not contradict one another. What Table 15 adds is that the combination is not less brittle along the axis equation (5) identifies: on one scorer and one set of texts the signed statistic falls from 0.988 to 0.517 on the self cell, the only cell that asks it its own question, while the supervised rows hold across the board. Liketropy and Fast-DetectGPT share one numerator: with the scorer as its own sampling model, the setting used here, the analytic criterion of Fast-DetectGPT is liketropy summed over the scored tokens and divided by its predicted standard deviation, which is why the two rows of Table 2 coincide. Across the six truncated cells, including the five cross-model cells for which Table 15 lists no liketropy value, their AUROCs differ by at most 0.0022, and under raw sampling by 0.006 to 0.015; the normalisation is the only difference between them. The framework's empirical content in this regime is therefore

Fast-DetectGPT's; what it adds is the absolute value and the theory, and the absolute value is what turns 0.988 into 0.110.

K LLM Selection Details

The six evaluators were chosen to vary the properties that could confound a result reported as common to all of them, since a failure that held only within one tokenizer family, one organisation's models or one parameter scale would not support the claims made here. Table 17 records the properties used.

Table 17: LLM selection. Sources: Falcon-RW-1B (Penedo et al., 2023), Llama-3-8B (Grattafiori et al., 2024), Gemma-7B (Gemma Team, 2024), Phi-2 (Javaheripi & Bubeck, 2023), OPT-1.3B (Zhang et al., 2022), Mistral-7B-v0.1 (Jiang et al., 2023). Architecture abbreviations: Tf, Transformer; MHA, multi-head attention; GQA, grouped-query attention (Ainslie et al., 2023); ALiBi, attention with linear biases (Press et al., 2022); RoPE, rotary position embeddings; SWA, sliding-window attention; FA, FlashAttention; GeGLU, GELU-gated linear unit. Multi denotes multilingual data.

Attribute	**Falcon-RW-1B**	**Llama-3-8B**	**Gemma-7B**
Organisation	TII (UAE)	Meta AI	Google DeepMind
Parameters	1B	8B	7B
Tokenizer Type	GPT-2 BPE	Custom tiktoken	SentencePiece
Vocabulary Size	50,257	128,256	256,000
Training Tokens	350B	15T	6T
Training Source	RefinedWeb	Web+Code+Multi	Web+Code+Math
Architecture	GPT-3+ALiBi+FA	Tf+GQA+RoPE	Tf+MHA+RoPE+GeGLU
Context Length	2,048	8,192	8,192

Attribute	**Phi-2**	**OPT-1.3B**	**Mistral-7B-v0.1**
Organisation	Microsoft Research	Meta AI	Mistral AI
Parameters	2.7B	1.3B	7B
Tokenizer Type	CodeGen BPE	GPT-2 BPE	SentencePiece
Vocabulary Size	51,200	50,272	32,000
Training Tokens	1.4T	180B	Undisclosed
Training Source	Synthetic textbook+Web	BookCorpus, Pile, Reddit	Undisclosed
Architecture	MixFormer	GPT-3 style	Tf+GQA+RoPE+SWA
Context Length	2,048	2,048	8,192

L Full Per-Dataset Results

Table 18: Condition A'' liketropy (original paper conditions), per dataset. $n = 400$ per dataset per model. The Human column follows the A'' scoring window, so it differs slightly from the Human column of Table 19. Values: mean $\pm$ std.

Dataset	Falcon-RW-1B		Gemma-7B		Llama-3-8B	
	A''	Human	A''	Human	A''	Human
RAID abstracts	0.653 ± 0.120	-0.039 ± 0.193	0.397 ± 0.104	-0.116 ± 0.183	0.427 ± 0.106	-0.073 ± 0.186
RAID books	0.681 ± 0.126	-0.015 ± 0.189	0.506 ± 0.151	-0.105 ± 0.221	0.517 ± 0.140	-0.057 ± 0.185
RAID news	0.569 ± 0.121	-0.021 ± 0.228	0.318 ± 0.116	-0.132 ± 0.221	0.394 ± 0.112	-0.055 ± 0.215
RAID poetry	0.659 ± 0.147	-0.061 ± 0.205	0.536 ± 0.144	-0.124 ± 0.189	0.516 ± 0.161	-0.094 ± 0.184
RAID recipes	0.360 ± 0.117	-0.061 ± 0.202	0.231 ± 0.118	-0.129 ± 0.155	0.259 ± 0.102	-0.068 ± 0.160
RAID reddit	0.566 ± 0.126	-0.033 ± 0.186	0.448 ± 0.118	-0.078 ± 0.191	0.470 ± 0.112	-0.055 ± 0.170
RAID reviews	0.664 ± 0.126	0.009 ± 0.208	0.509 ± 0.125	-0.056 ± 0.207	0.533 ± 0.118	-0.011 ± 0.194
XSum	0.560 ± 0.124	0.033 ± 0.145	0.348 ± 0.109	-0.067 ± 0.130	0.410 ± 0.113	0.036 ± 0.127
Average	$\mathbf{0.589 \pm 0.160}$	$\mathbf{-0.024 \pm 0.198}$	$\mathbf{0.412 \pm 0.160}$	$\mathbf{-0.101 \pm 0.192}$	$\mathbf{0.441 \pm 0.149}$	$\mathbf{-0.047 \pm 0.183}$

Dataset	Phi-2		OPT-1.3B		Mistral-7B-v0.1	
	A''	Human	A''	Human	A''	Human
RAID abstracts	0.427 ± 0.202	-0.140 ± 0.200	0.647 ± 0.138	-0.150 ± 0.216	0.395 ± 0.093	-0.072 ± 0.167
RAID books	0.518 ± 0.184	-0.117 ± 0.207	0.669 ± 0.145	-0.155 ± 0.208	0.466 ± 0.117	-0.071 ± 0.169
RAID news	0.486 ± 0.140	-0.105 ± 0.227	0.522 ± 0.127	-0.166 ± 0.253	0.333 ± 0.102	-0.087 ± 0.201
RAID poetry	0.501 ± 0.203	-0.232 ± 0.223	0.571 ± 0.165	-0.317 ± 0.241	0.469 ± 0.124	-0.146 ± 0.182
RAID recipes	0.264 ± 0.127	-0.232 ± 0.206	0.405 ± 0.113	-0.175 ± 0.223	0.166 ± 0.081	-0.118 ± 0.148
RAID reddit	0.490 ± 0.143	-0.120 ± 0.184	0.531 ± 0.122	-0.191 ± 0.219	0.414 ± 0.108	-0.072 ± 0.166
RAID reviews	0.540 ± 0.199	-0.097 ± 0.221	0.644 ± 0.129	-0.143 ± 0.242	0.464 ± 0.114	-0.047 ± 0.183
XSum	0.485 ± 0.145	-0.060 ± 0.154	0.526 ± 0.122	-0.064 ± 0.148	0.344 ± 0.097	-0.005 ± 0.119
Average	$\mathbf{0.464 \pm 0.189}$	$\mathbf{-0.138 \pm 0.212}$	$\mathbf{0.564 \pm 0.157}$	$\mathbf{-0.170 \pm 0.230}$	$\mathbf{0.381 \pm 0.142}$	$\mathbf{-0.077 \pm 0.173}$

Table 19: Condition A liketropy (raw sampling, first 30 tokens excluded from scoring), per dataset. $n = 400$ per dataset per model. Values: mean $\pm$ std.

Dataset	Falcon-RW-1B		Gemma-7B		Llama-3-8B	
	A	Human	A	Human	A	Human
RAID abstracts	-0.022 ± 0.166	-0.034 ± 0.207	-0.058 ± 0.169	-0.089 ± 0.192	-0.013 ± 0.154	-0.049 ± 0.185
RAID books	-0.015 ± 0.175	-0.014 ± 0.199	-0.088 ± 0.204	-0.083 ± 0.209	-0.022 ± 0.166	-0.039 ± 0.200
RAID news	-0.029 ± 0.164	-0.026 ± 0.233	-0.034 ± 0.152	-0.108 ± 0.224	-0.013 ± 0.161	-0.048 ± 0.219
RAID poetry	-0.030 ± 0.178	-0.008 ± 0.217	-0.120 ± 0.235	-0.058 ± 0.196	-0.024 ± 0.185	-0.040 ± 0.196
RAID recipes	-0.026 ± 0.174	-0.032 ± 0.205	-0.073 ± 0.175	-0.093 ± 0.153	-0.011 ± 0.138	-0.045 ± 0.152
RAID reddit	-0.034 ± 0.177	-0.027 ± 0.199	-0.068 ± 0.192	-0.069 ± 0.185	-0.026 ± 0.165	-0.048 ± 0.179
RAID reviews	-0.011 ± 0.176	0.002 ± 0.219	-0.058 ± 0.184	-0.047 ± 0.211	-0.025 ± 0.159	-0.007 ± 0.207
XSum	-0.022 ± 0.183	0.052 ± 0.156	-0.053 ± 0.157	-0.055 ± 0.137	-0.033 ± 0.162	0.044 ± 0.132
Average	$\mathbf{-0.024 \pm 0.174}$	$\mathbf{-0.011 \pm 0.207}$	$\mathbf{-0.069 \pm 0.187}$	$\mathbf{-0.075 \pm 0.191}$	$\mathbf{-0.021 \pm 0.162}$	$\mathbf{-0.029 \pm 0.188}$

Dataset	Phi-2		OPT-1.3B		Mistral-7B-v0.1	
	A	Human	A	Human	A	Human
RAID abstracts	-0.008 ± 0.145	-0.096 ± 0.205	-0.026 ± 0.165	-0.121 ± 0.225	0.000 ± 0.147	-0.044 ± 0.172
RAID books	-0.022 ± 0.170	-0.085 ± 0.219	-0.020 ± 0.185	-0.155 ± 0.215	-0.004 ± 0.149	-0.055 ± 0.182
RAID news	-0.016 ± 0.163	-0.086 ± 0.231	-0.007 ± 0.168	-0.159 ± 0.258	-0.009 ± 0.148	-0.065 ± 0.201
RAID poetry	-0.030 ± 0.191	-0.120 ± 0.227	-0.045 ± 0.172	-0.259 ± 0.238	-0.012 ± 0.163	-0.062 ± 0.194
RAID recipes	-0.008 ± 0.158	-0.141 ± 0.208	-0.019 ± 0.158	-0.135 ± 0.229	-0.007 ± 0.125	-0.069 ± 0.146
RAID reddit	-0.021 ± 0.184	-0.084 ± 0.196	-0.020 ± 0.173	-0.212 ± 0.226	-0.006 ± 0.143	-0.058 ± 0.172
RAID reviews	-0.010 ± 0.179	-0.063 ± 0.236	-0.020 ± 0.172	-0.151 ± 0.262	-0.001 ± 0.150	-0.033 ± 0.194
XSum	-0.006 ± 0.165	-0.023 ± 0.166	-0.020 ± 0.152	-0.050 ± 0.154	-0.005 ± 0.139	0.010 ± 0.126
Average	$\mathbf{-0.015 \pm 0.170}$	$\mathbf{-0.087 \pm 0.215}$	$\mathbf{-0.022 \pm 0.169}$	$\mathbf{-0.155 \pm 0.235}$	$\mathbf{-0.005 \pm 0.146}$	$\mathbf{-0.047 \pm 0.177}$

In Table 20, recipes is the extreme row under B'', in the direction the mechanism predicts: it is the lowest entry for every evaluator, between 0.037 and 0.174 against 0.155 to 0.478 elsewhere, because a formulaic domain leaves little entropy for truncation to remove. This is where most of the 31 exceptions in Table 10 arise (16 on recipes alone).

Table 20: Cross-model liketropy per dataset, conditions B (raw) and B'' (truncated). Each entry is the evaluator scoring the five other base models' text, pooled over those five generators; $n = 400$ per generator per dataset, so 2,000 per entry. Values: mean ± std.

Dataset	Falcon-RW-1B	OPT-1.3B	Phi-2	Gemma-7B	Llama-3-8B	Mistral-7B-v0.1
Condition B: a different model, raw full-vocabulary sampling						
RAID abstracts	-0.184 ± 0.224	-0.294 ± 0.250	-0.287 ± 0.227	-0.330 ± 0.244	-0.327 ± 0.233	-0.272 ± 0.198
RAID books	-0.177 ± 0.225	-0.374 ± 0.258	-0.300 ± 0.231	-0.312 ± 0.234	-0.342 ± 0.232	-0.274 ± 0.193
RAID news	-0.155 ± 0.220	-0.382 ± 0.262	-0.255 ± 0.233	-0.339 ± 0.226	-0.362 ± 0.220	-0.282 ± 0.185
RAID poetry	-0.161 ± 0.245	-0.476 ± 0.287	-0.270 ± 0.260	-0.285 ± 0.259	-0.325 ± 0.244	-0.287 ± 0.215
RAID recipes	-0.179 ± 0.223	-0.329 ± 0.294	-0.275 ± 0.235	-0.279 ± 0.217	-0.292 ± 0.214	-0.264 ± 0.171
RAID reddit	-0.143 ± 0.223	-0.409 ± 0.268	-0.191 ± 0.222	-0.262 ± 0.237	-0.264 ± 0.228	-0.259 ± 0.201
RAID reviews	-0.179 ± 0.227	-0.391 ± 0.263	-0.230 ± 0.226	-0.288 ± 0.236	-0.309 ± 0.223	-0.266 ± 0.186
XSum	-0.161 ± 0.219	-0.395 ± 0.278	-0.269 ± 0.234	-0.362 ± 0.224	-0.357 ± 0.217	-0.288 ± 0.177
Average	$\mathbf{-0.167 \pm 0.226}$	$\mathbf{-0.381 \pm 0.275}$	$\mathbf{-0.260 \pm 0.236}$	$\mathbf{-0.307 \pm 0.237}$	$\mathbf{-0.322 \pm 0.229}$	$\mathbf{-0.274 \pm 0.192}$
Condition B'': a different model, top-k 40 and top-p 0.96						
RAID abstracts	0.389 ± 0.155	0.298 ± 0.164	0.276 ± 0.156	0.245 ± 0.141	0.238 ± 0.139	0.206 ± 0.128
RAID books	0.475 ± 0.162	0.365 ± 0.168	0.329 ± 0.176	0.316 ± 0.176	0.287 ± 0.161	0.266 ± 0.135
RAID news	0.363 ± 0.149	0.223 ± 0.154	0.260 ± 0.148	0.163 ± 0.145	0.164 ± 0.144	0.155 ± 0.122
RAID poetry	0.478 ± 0.166	0.290 ± 0.169	0.345 ± 0.185	0.353 ± 0.169	0.322 ± 0.164	0.275 ± 0.146
RAID recipes	0.174 ± 0.130	0.127 ± 0.136	0.059 ± 0.147	0.077 ± 0.129	0.082 ± 0.116	0.037 ± 0.098
RAID reddit	0.412 ± 0.136	0.286 ± 0.141	0.323 ± 0.140	0.307 ± 0.140	0.301 ± 0.130	0.248 ± 0.124
RAID reviews	0.474 ± 0.159	0.361 ± 0.170	0.377 ± 0.183	0.356 ± 0.163	0.317 ± 0.161	0.261 ± 0.143
XSum	0.352 ± 0.148	0.223 ± 0.158	0.256 ± 0.149	0.161 ± 0.143	0.172 ± 0.139	0.157 ± 0.119
Average	$\mathbf{0.390 \pm 0.178}$	$\mathbf{0.272 \pm 0.174}$	$\mathbf{0.278 \pm 0.186}$	$\mathbf{0.247 \pm 0.180}$	$\mathbf{0.235 \pm 0.167}$	$\mathbf{0.201 \pm 0.149}$

M Per-Model Liketropy by Scored Text Length

Section 4.2.2 reports the length sweep in summary. In full: self-liketropy holds between -0.020 and -0.035 across the five lengths while its pooled standard deviation falls from 0.179 to 0.096, so the centre does not move and only the spread does, which is the variance reduction Theorem 1(a) predicts. Human text tightens in step, from 0.175 to 0.109. Averaged over the six evaluators, human text sits 0.014 nats further from zero than the model's own output at 200 scored tokens and 0.011 further at 800; that difference never exceeds 0.018 at any length and carries no monotone trend.

Table 21: Each evaluator's liketropy on its own raw-sampled output and on human text, at each scored text length. Values: mean ± std. The pooled block combines the six evaluators, its standard deviation adding the variance of the cell means to the within-cell variance.

Length	Falcon-RW-1B		Gemma-7B		Llama-3-8B	
	A	Human	A	Human	A	Human
200	-0.0073 ± 0.1875	0.0367 ± 0.1767	-0.0780 ± 0.1958	-0.0472 ± 0.1431	-0.0054 ± 0.1605	0.0177 ± 0.1565
300	-0.0289 ± 0.1524	0.0470 ± 0.1659	-0.0714 ± 0.1305	-0.0399 ± 0.1268	-0.0064 ± 0.1351	0.0287 ± 0.1284
400	-0.0334 ± 0.1340	0.0652 ± 0.1566	-0.0569 ± 0.1454	-0.0400 ± 0.1130	-0.0317 ± 0.1155	0.0217 ± 0.1100
600	-0.0272 ± 0.1060	0.0838 ± 0.1384	-0.0795 ± 0.1171	-0.0299 ± 0.0836	-0.0384 ± 0.0958	0.0233 ± 0.0735
800	-0.0097 ± 0.0842	0.0909 ± 0.1384	-0.0947 ± 0.1244	-0.0319 ± 0.0625	-0.0131 ± 0.0811	0.0234 ± 0.0580

Length	Phi-2		OPT-1.3B		Mistral-7B-v0.1	
	A	Human	A	Human	A	Human
200	0.0008 ± 0.1696	-0.0405 ± 0.1911	-0.0522 ± 0.1923	-0.0652 ± 0.1851	-0.0146 ± 0.1507	-0.0328 ± 0.1674
300	-0.0052 ± 0.1234	-0.0288 ± 0.1768	0.0019 ± 0.1357	-0.0579 ± 0.1691	-0.0075 ± 0.1212	-0.0255 ± 0.1452
400	-0.0268 ± 0.1191	-0.0164 ± 0.1492	-0.0182 ± 0.1268	-0.0660 ± 0.1561	-0.0133 ± 0.1085	-0.0079 ± 0.1262
600	-0.0403 ± 0.1025	-0.0086 ± 0.1129	-0.0196 ± 0.0860	-0.0749 ± 0.1054	-0.0075 ± 0.0859	0.0056 ± 0.1087
800	-0.0305 ± 0.0797	-0.0054 ± 0.1031	-0.0209 ± 0.0884	-0.0837 ± 0.0918	-0.0116 ± 0.0806	0.0128 ± 0.0919

Pooled over the six evaluators

Length	A	Human
200	-0.0261 ± 0.1792	-0.0219 ± 0.1746
300	-0.0196 ± 0.1358	-0.0127 ± 0.1579
400	-0.0300 ± 0.1262	-0.0072 ± 0.1430
600	-0.0354 ± 0.1020	-0.0001 ± 0.1165
800	-0.0301 ± 0.0958	$+0.0010 \pm 0.1087$

N SAMPLE CONTINUATIONS

This appendix presents sample continuations from Falcon-RW-1B and Llama-3-8B at each top-k level, each conditioned on a single human-written prompt drawn from the XSum dataset. Each depth was sampled independently, so the continuations diverge from one another almost immediately; what a block holds fixed is the prompt, not the sampling path. These generations come from a companion sweep on XSum of $n = 400$ documents rather than from the instrumented run of Appendix E, so the values here are comparable in trend but not in level with Figure 2.

The samples illustrate the qualitative pattern discussed in Section 4.1. Liketropy falls steadily as the truncation depth increases, pooling at $+0.874$ at $k = 5$ and -0.027 under full-vocabulary sampling for Falcon-RW-1B, while the text itself remains fluent across the truncated range. In the documents shown the descent is not strictly monotone: Llama-3-8B document 311 places $k = 640$ above $k = 320$, which is per-document sampling noise rather than a departure from the pooled trend. Under full-vocabulary (*RAW*) sampling the smaller model degrades visibly even as its liketropy approaches zero, which is the gap between fluency and score that the argument turns on.

FALCON-RW-1B, XSUM DOCUMENT 170

Shared prompt: More than 1.7 million Npower customers could save £261 a year, simply by moving to another Npower tariff. As many as

Human liketropy $= +0.188$

6.6 million British Gas customers could save £129 a year, according to Ofgem. However British Gas says it has changed its prices, so the saving is now £43. The figures also reveal that SSE has almost all of its customers on the most expensive form of tariff. 91% of SSE customers are on standard variable tariffs, compared to just 9% at First Utility. The number of people stuck on standard tariffs is controversial, as most fixed-rate deals are cheaper. In the UK as a whole, 66% of households are on standard tariffs. However Npower said that customers were unlikely to be able to save £261, as fixed-price deals were usually limited to a certain number of customers. Once a certain number of people have signed up, the deal is withdrawn. Last month, the business and energy secretary, Greg Clark, met energy suppliers, amid claims that some of them were making large profits from customers who are on standard variable tariffs. The industry denied

$k = 5$ liketropy $= +0.981$

6.3 million Npower customers will be able to save £1.2m a year if their current Npower tariff is upgraded. Npower's latest tariff offers an average saving of £261 per year. The average customer could save an additional £1,200 a year by switching to Npower's new standard tariff. Npower is also making a significant investment in the UK's energy infrastructure and will be able to offer customers a wider choice of products and services, including smart meters, which will give them more control over their energy usage. The new Npower standard tariff has been developed in response to consumer feedback and feedback from industry experts. The new Npower standard price includes a number of features that Npower customers currently don't have access to, such as a new smart meter, online bill checking and access to the Npower app. Npower is also introducing a new energy-efficient home insulation scheme, which will be available to all

$k = 10$ liketropy $= +0.770$

6.3 million Npower customers would save £944, while a further 2.5 million would save £827. The average household would save £10,000 a year. Npower customers could be saving on average £10,000 a year The findings, from analysis by the think tank IPPR, show that the Npower Energy Switch could be saving 1.2 million homes a year in the UK and 2.5 million homes worldwide. Npower's chief executive, Ian Marchant, recently admitted that he was 'not sure' what customers would save as the company was 'still working out all the details'. But IPPR's analysis shows that a simple switching from Npower's standard tariff to an alternative would save customers up to £2,400. The switch is a simple way for customers to save money without having to make a huge change to their energy habits. It's also a much more cost effective

$k = 20$ liketropy $= +0.575$

6.3 million Npower customers would save £944,988,876 by switching. Ofgem has announced that customers of Npower's unregulated electricity market could save an average of £1,719 a year by signing up to one of the new tariff options. The average saving, taking the lowest-priced option, is £1,867 a year or £261 a month. As many as 6.3 million Npower customers could save £944,988,876 by switching to another Npower tariff. Ofgem's report found that more than 1.7 million customers – more than a third - could save between 1% and 6.3% on their electricity bills. Npower has said that it has not changed the tariff for anyone at this time and is in discussions with Ofgem to resolve any disputes. However, Ofgem is calling for any customer with a complaint on the market to report it.

$k = 40$ liketropy $= +0.432$
6.3 million could have their bills cut by as much as £1,000 a year if they simply took advantage of our cheaper prices and fixed rate tariffs. Npower is also increasing the amount they pay for your electricity. The average customer could save an additional £80 per year. The news comes as Npower, who have been criticised for paying thousands of its customers below the regulated electricity market rate through their tariffs, are set to be charged £3.3m for the energy they've bought for the UK. How do I know if my electricity is costing me more or less than a Fixed or Variable Npower tariff? You can find out if electricity is costing you more or less by comparing it to Npower deals. The chart shows how Npower deals can help you save money on your monthly gas and electricity bills. How do I know if more people are moving to Npower, which will mean bigger benefits for me? Npower customers can get a free no

$k = 80$ liketropy $= +0.345$
6.3 million could have their next electricity bill reduced by £250 a year, based on the latest winter energy price forecast. Other customers could save by switching to Npower's new Flexible Npower tariff. Npower has also slashed the energy efficiency package prices, especially those customers who are on a cheaper prepayment meter. All new Npower customers on prepayment meters can enjoy £150 off the Energy & Fuel Saving Programme, reducing the overall cost to £84 for those who top up for the winter season. Customers only have to apply and open their pre-payments at a participating store to get the discount. They do not need to re-top up if they are on prepayment rate. However, Npower warned that its winter price forecast for the lowest-cost tariff as well as the new Flexible tariff and its new energy saving package prices will only apply to new prepayment and prepayment top-up customers. Existing prepayment

$k = 160$ liketropy $= +0.241$
6.3 million Npower customers would save £948,988! If you change to another Npower tariff, you could immediately save yourself a shed load of money. This is my Top Npower Switch Tip. If you have an Npower gas package, I am quite sure you are on a cheaper gas tariff that is way too expensive for you! If your Npower gas bill is in the thousands and doesn't get reduced because you are on broadband, then you are grossly overpriced! You don't know how much you're paying because the Npower accountants don't give out information! The average electricity bill for an Npower customer is roughly £150 in June, but Npower don't give out any information. You need to self-serve your online meter account and you'll see that your current Npower tariff is £0.00 Npower's gas tariffs are no better.

$k = 320$ liketropy $= +0.150$
6.3 million Npower customers would save £948,988! If you change to another Npower tariff, you could immediately save yourself a shed load of money. This is my Top Npower Switch Tip. If you have an Npower gas package, I am quite pleased to announce you might be able to save up to £3,688! If your Npower package includes electric in your bill, you can save even more money. With a great broadband change from £5.75 a month, don't forget to save even more by switching to another Npower dual fuel or fibre product. If you do decide to change to another Npower tariff, you may be able to immediately save on your electricity bill. A recent Npower customer saving found that based on the tariff you selected at end 2005 and the supply you choose, you would save £4,566, which is a shed load of money! If you do not have an Npower

$k = 640$ liketropy $= +0.083$
6.3 million Npower customers would save £948,988! If you change to another Npower tariff, you could immediately save yourself a shed load of money. This is my Top Npower Switch Tip. If you have an Npower gas bill, I am sorry to give you bad news, but I'm about to tell you your situation is completely mind-boggling. If you are a Npower Gas customer, chances are, you probably have a number of Npower gas products you might need to consider to get the best deal from Natwest. There's a fair amount you might not know. Trust me, I'm not just saying this: I discovered how complicated it all is The first time I applied for Natwest. It was for my first mortgage four years ago. All thanks to me, you might how to save several hundred pounds, which will help to fund any plans you have. I came across this video on

RAW liketropy $= +0.080$
6.3 million Npower customers would save £948,988! If you change to another Npower tariff, you could immediately save yourself a shed load of money. This is my Top Npower Switch Community Rate Saving Mistakes. Here's my top tips, especially those that give you the biggest performance/savings. Npower is repeatedly asked by customers for help on refunds through their website. Occasionally they reply and offer a few pointers; however, all they change is the web address, cannot offer any assistance, and only confirm subject to hoards of hoops you have to jump through to get a refund. Npower do not offer refunds on electricity bills. The current regulations make it illegal to rate adjust on your electricity bill. Nevertheless, Npower have been found to have breeched the current legislation. This would mean that you could be charged more than double without compensating you, which is misconduct and is against the law by any fair billing authority. Help is

LLAMA-3-8B, XSUM DOCUMENT 311

Shared prompt: The supermarket's like-for-like sales, which strip out new stores, fell by 0.5% in the period to 11 March.

Human liketropy $= +0.087$

However, Argos sales rose 4.3%, resulting in a 0.3% increase excluding fuel across the Sainsbury's group. Sainsbury's said the market remained "very competitive" and the impact of cost pressures "remains uncertain". The supermarket's performance was worse than the 0.1% rise in like-for-like sales seen in the 15 weeks to 7 January. Sainsbury's convenience stores saw total sales growth of almost 7%, with 10 new outlets opening in the quarter. Chief executive Mike Coupe said customers had welcomed Sainsbury's "differentiated food offer", with new products including butternut squash waffles and sweet potato tagliatelle. General merchandise sales were down 4%, which Sainsbury's said was due to Easter and Mother's Day falling later this year. Sainsbury's bought catalogue retailer Argos last year as part of its £1.4bn takeover of Home Retail Group. The retailer has opened another 11 Argos Digital stores in Sainsbury's supermarkets since the start of the year, bringing the total to 41, while there are now eight

$k = 5$ liketropy $= +0.754$

That compares with a 0.2% increase in the fourth quarter. The group said the sales performance was impacted by the timing of Easter, which fell in April last year, and by a "challenging" comparison with the previous year, when it introduced its "Price Crunching" initiative. However, Tesco's online sales were up almost 20% during the quarter. It said its non-food sales grew by 5%, while general merchandise rose 3.5%. Tesco said it was on course to deliver £500m of savings this year, and that the cost of implementing the turnaround plan was "on track". Tesco said it had made progress with its turnaround plan in the UK, including the introduction of a new loyalty card scheme, Clubcard Plus. It has 10.5m customers on the card. It has also made improvements to its online business, including the launch of the "click and collect" service in 1,400 stores.

$k = 10$ liketropy $= +0.520$

That compared with a 0.2% increase in its fourth quarter. The firm said the sales decline was driven by a "challenging" environment on the UK high street, with a "marked shift" in the balance of power towards suppliers. It added that the cost of living squeeze was having a "significant impact" on consumer spending. The retailer's online sales continued to grow strongly, up 10.8% in the quarter, but this was at a lower rate than in the previous quarter. Asda has 631 supermarkets and other stores and employs 161,600 people. It is the second biggest supermarket in the UK, behind Tesco. In the three months to 11 March Asda made sales of £4.3bn.

$k = 20$ liketropy $= +0.459$

That compared with a 0.2% increase in its fourth quarter. The group posted a total sales rise of 4.7% in the period. Morrisons has come under severe pricing pressure from the "Big Four" supermarket rivals, which it blames on the "big four" retailers. As a result, it has announced another round of store closures. Chief executive David Potts said the latest fall in sales was mainly in the "Big Four" rivals' core areas of fresh food. Morrisons has been working to move its focus to the fresh products it has excelled at in the past and away from the low-margin processed convenience foods it says are driving prices downwards. Mr Potts said the group had seen strong growth in its convenience format stores, especially in the north and midlands. He said food sales had risen strongly in its wholesale arm, which supplies food to other convenience stores, and that the group was now seeing growth in its core online grocery business. In the

$k = 40$ liketropy $= +0.343$

That compared with a 0.2% increase in its fourth quarter. The rise in the pound following the EU referendum result hit Asda's performance, along with food price inflation, which is at a five-year high of 4.2%. The grocery retailer also reported a slump in sales from its George clothing range, down 11% compared with the same period last year. Asda said this was due to less favourable product mix and a "more promotional" environment. Sales of branded goods grew 1.4%, while own-branded products fell by 1.8%, it said in a statement. Pre-tax profits were impacted by more than £100m due to one-off costs, including for an overhaul of its distribution network and a reduction in the value of its pension assets. It said it expected a further reduction of that value in the current first quarter. Asda said the value of its pension assets was cut because of the recent decline in gilt yields. However, Asda

$k = 80$ liketropy $= +0.221$

That compared with a 0.2% increase in its fourth quarter. The rise in the pound following the EU referendum result has hurt profits while the business suffers inflationary increases in food prices, it said. Morrisons' shares dropped by nearly 11%, at 113.15p, down by 13p, while the FTSE 100 index was down by 0.33% at 6938. But David Potts, CEO of Morrisons, told BBC Radio 5 live's Wake Up to Money: "We are doing well, we are out-performing the market and the group is trading well." 'Customers have seen price increases' Profits fell by

50% for the year on Wednesday from a year earlier to £178m. This result fell below analysts' expectations, which had been expecting an annual figure of £204m. The company said the fall in profits was due to the "mix of factors", including a "deterioration in

$k = 160$ liketropy $= +0.207$

That compared with a 0.2% increase in its store sales. The rise in the pound following the EU referendum result has hurt profits while the business suffers inflationary increases in food prices, input costs and transport costs. The chain is also trying to reduce its costs by £500 million over the next five years by cutting its distribution costs and its net property costs. The group managed to increase its operating profit by 11% in the half but warned on the 2018 results. Chief executive David Potts said: "Trading since the end of the year has been difficult. Asda is in the process of being bought by US retail giant Walmart after the group dropped its takeover of Sainsbury's for an undisclosed amount in February. The deal is expected to be approved by the Competition and Markets Authority later this year.

$k = 320$ liketropy $= +0.044$

That compared with a 0.2% increase in its store sales. The rise in the pound following the EU referendum result has hurt profits while the reallocation of Tesco's own-brand pricing strategy, as well as the launch of its Jack's brand, has also begun to have an impact. UK like-for-like sales were down by 2.5%, dipping below expectations. Meanwhile, the wider group remained in the black in the third quarter, helped by cost cutting. Sales rose 1%. Like-for-like sales in 29 international markets where the group operates were up 3.6%. The business was formerly the Tesco Lotus convenience chain of shops in Thailand, having been bought by CP Group for an undisclosed amount in 2013. CP Foods, the largest listed company and agro-industrial conglomerate in Thailand, has agreed to buy a 51% stake in the consortium, with the remaining assets divided equally between SSI, Samudera and MC Group.

$k = 640$ liketropy $= +0.111$

That compared with a 0.2% increase in its store sales. The rise in the pound following the EU referendum result has hurt profits while the reallocation of spending into lower inflation goods is proving a barrier for the retailer. Morrisons, which is aiming to complete the sale of 140 of its stores by the end of April to form a new company run by its former chief executive, David Potts, said it has sold 109 of its stores. With debts of £130m, the retailer has offered bidders a chance to buy the remaining stores in a new round of auctions in April and May. Morrisons' shares have fallen around 50% since the vote on 23 June, more than twice the fall suffered by the FTSE 100 so that the price of £2.32 has slipped to an all-time low this week. Morrisons said the value of its remaining assets was "subject to significant uncertainty", but a new chief executive is set to

RAW liketropy $= -0.079$

Marks & Spencer's (M&S) did better, with a 0.2% rise in a sector in which supermarkets are cannibalising traditional High Street stores. Morrisons's sales were bolstered by online posting up 46.5%, while Baskin-Robbins franchisees kicked in 11.6% growth. New chief executive David Potts took over at the supermarket at the start of February from wafer-owner Marc Bolland. Potts said today: "Since joining Morrison's last year, I have been amazed by the quality and dedication shown by our staff, I have been truly humbled by their customary normal humility. "That has enabled me to quickly identify the key issues affecting our performance and I have been delighted to see the progress we made in the first step of our turnaround plan." He continued: "I am determined that we can achieve the ambitions set out in our strategy, absolutely all delivery partners have to be aligned and motivated, including me." A £