No Model Required: Text Entropy Rate Filtering Mitigates Iterative Fine-Tuning Collapse

Lewis Mitchell
Adelaide Data Science Centre
School of Mathematical Sciences
Adelaide University
Adelaide SA 5005, Australia
lewis.mitchell@adelaide.edu.au

Abstract

Iterative fine-tuning on synthetic data causes *model collapse*: output diversity narrows as rare patterns are progressively lost, a signature most visible as phrase-level repetition. Existing mitigations either require model log-probabilities, an external oracle, or continued access to real human data. Here we develop a new approach grounded in mathematical information theory: the non-parametric Kontoyiannis entropy rate estimator $\hat{H}_K$, computed entirely from raw text via match-length statistics, with no model of any kind. We show that this is in fact a *superior* training-data filter on text-diversity metrics in a fully-synthetic, single-lineage fine-tuning setting. In a six-generation QLoRA collapse experiment on Llama-3.1-8B, logprob-based filtering (the most established model-access-requiring baseline) provides no significant text-diversity benefit on any metric ($p > 0.23$), whereas $\hat{H}_K$-filtering yields $+42\%$ unique trigrams, $+30\%$ vocabulary, and -19% repetition (all $p < 0.001$). We validate $\hat{H}_K$ as a cross-domain entropy proxy ($\beta = 0.924$, $R^2 = 0.746$) and collapse detector ($\rho = +0.454$, $p < 0.0001$) across 4 domains, 2 temperatures, 2 generator–scorer model pairs, and 1,520 generated documents. Our results demonstrate that information theoretic approaches to collapse mitigation are efficient, and suggest new approaches for maintaining multi-agent diversity.

1 Introduction

Large language models trained on their own outputs face a self-reinforcing degradation known as *model collapse*: when a model's outputs are used to fine-tune successor models, the entropy of the output distribution declines across generations, diversity narrows, and rare linguistic patterns are progressively lost [Shumailov et al., 2024, Alemohammad et al., 2024]. As synthetic content becomes an increasing fraction of web-scraped training corpora, collapse mitigation is shifting from a theoretical concern to an engineering requirement for production training pipelines.

Existing mitigations for model collapse fall into one of three categories: (A) model logprob filters such as *surplexity filtering* [Gambetta et al., 2026], which scores documents by the model's own per-token Shannon entropy and retains those the model finds surprising; (B) external oracle or classifier approaches utilising an independent language model; and (C) data accumulation strategies that re-inject real human data at each generation [Gerstgrasser et al., 2024, Bertrand et al., 2024]. Among filtering approaches that do not require the injection of real, expensive human data, surplexity filtering is the most established model-access-requiring baseline. Gambetta et al. [2026] demonstrate that it delays collapse as measured by the model's own entropy and perplexity scores, but its effect on independent text-diversity measures has not been evaluated. Throughout this paper we

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

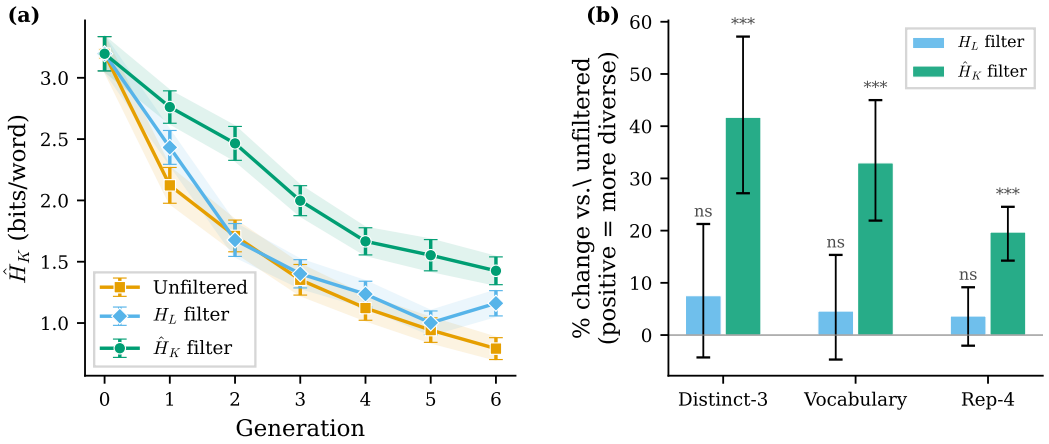


Figure 1: Filtering on text-only entropy preserves diversity; logprob-based filtering does not. **(a)** Kontoyiannis entropy rate $\hat{H}_K$ across six fine-tuning generations of Llama-3.1-8B-Instruct under three training-data conditions: unfiltered (rapid collapse), H_L-filtered (intermediate), and $\hat{H}_K$-filtered (slowest decline). Shaded bands: ± 1 SEM ($n = 80$ documents per generation). **(b)** Percentage change relative to the unfiltered baseline at generation 6 for three text-diversity metrics (Distinct-3, Vocabulary size, Rep-4). Error bars: bootstrap 95% CIs. $\hat{H}_K$-based filtering yields highly significant improvements on all three metrics ($^{***}p < 0.001$), whereas H_L-based filtering is non-significant on all three.

compare against mean per-token Shannon entropy H_L specifically, which is closely related to but not necessarily identical to surplexity: both are monotonically increasing functions of the model's output entropy, but surplexity uses the sample surprise of the generated tokens whereas H_L averages over the full output distribution, so the two induce highly similar but not necessarily identical document rankings. Moreover, Guo et al. [2024] find that linguistic acceptability filtering can actually *worsen* collapse, underscoring that the choice of filtering criterion is consequential.

Surplexity filtering requires a forward pass through the model being trained, with access to the full next-token distribution at each position. Such access is unavailable for closed-source APIs, computationally expensive at scale, and version-dependent – the same document receives different surplexity scores as the model is updated. Data accumulation avoids model access but presupposes a continuing supply of human-authored text, which is costly and slow to generate, and unavailable in a fully synthetic closed loop. A text-only collapse filter that requires no model of any kind would be strictly preferable, even if it were found to be less effective than surplexity filtering.

Motivated by results from mathematical information theory, we propose the **Kontoyiannis entropy rate estimator** $\hat{H}_K$ [Kontoyiannis et al., 1998] as a text-only collapse filter and show that it outperforms logprob-based filtering. $\hat{H}_K$ estimates the entropy rate of a symbolic sequence by measuring match lengths: at each position, it records the length of the longest prefix of the remaining text that has previously appeared in the document. Long matches signal repetition in the text; short matches signal novelty. The Kontoyiannis estimator has previously been used to quantify information flow in social media text [Bagrow et al., 2019, Bagrow and Mitchell, 2018, Pond et al., 2020] and requires no model, no API, and no GPU. To our knowledge, no prior work has applied $\hat{H}_K$ or any compression-rate statistic as a collapse filter; it is the only proposed filter of this type requiring no model of any kind – not a forward pass, not an oracle, not a reference classifier.

Figure 1 previews our central result. In a six-generation QLoRA fine-tuning experiment on Llama-3.1-8B-Instruct [Dubey et al., 2024], each generation's model is fine-tuned from the previous one – a continuous model lineage reflecting the realistic deployed scenario – and the filter operates in a fully synthetic closed loop with no human data available for re-injection. H_L-based training-data filtering – the most established model-access-requiring baseline – provides no significant text-diversity benefit on any measured metric ($p > 0.23$), whereas **$\hat{H}_K$-based filtering yields highly significant improvements** ($p < 0.001$) on Distinct-3, vocabulary size, and repetition rate – despite $\hat{H}_K$ requiring

no model access at all. The two filters select substantially different training documents (mean within-domain Jaccard similarity ≈ 0.35), indicating that $\hat{H}_K$ is not merely approximating surplexity but operating on a complementary axis: resistance to phrase-level repetition, the primary surface manifestation of fine-tuning collapse.

We validate $\hat{H}_K$ through three experiments: establishing it as a proxy for logprob entropy across domains, temperatures, and model pairs (Section 3); characterising when it does and does not track collapse (Section 4); and demonstrating that $\hat{H}_K$-based filtering outperforms logprob-based filtering in a controlled iterative fine-tuning experiment (Section 5). We release our code and data for reproducibility[1].

2 Background and Related Work

Model collapse. Shumailov et al. [2024] demonstrate that iterative training on model-generated data causes progressive loss of distributional tails: output variance shrinks across generations as the model reinforces its existing biases. The phenomenon is formalised by Alemohammad et al. [2024], who show that without real-data anchoring, iterative self-distillation inevitably converges to low-entropy output. Briesch et al. [2023] provide the first LLM-specific empirical study of self-consuming training loops, finding that diversity declines in proportion to the synthetic data fraction. Measuring collapse in terms of *text-level* diversity is a more recent development: Guo et al. [2024] conduct iterative fine-tuning on NLG tasks and track lexical, syntactic, and semantic diversity across generations, observing that syntactic diversity deteriorates fastest and creative tasks are most vulnerable. Both works characterise collapse through output distribution statistics; we measure it directly via the entropy rate of the generated text. Analogous diversity collapse has been documented in multi-agent LLM systems, where structural coupling during iterative interaction drives premature semantic convergence even without any weight update [Chen et al., 2026]; the present work identifies a complementary mechanism operating through the fine-tuning data loop.

Mitigating model collapse. Existing mitigations fall into three broad categories, each imposing a distinct requirement on the practitioner.

(A) Model-logprob methods. Gambetta et al. [2026] propose *surplexity* – the product of token-level surprise and perplexity – as a data-quality signal for collapse mitigation. In practice, retaining high-surplexity documents is closely related to filtering by the mean Shannon entropy of the generator's next-token distribution, which we denote H_L. This is the text-level analogue of perplexity-based data filtering widely used in large-scale pretraining pipelines [Wenzek et al., 2020]. Zhu et al. [2025] resample tokens whose predicted probability exceeds a threshold from a prior language model, requiring token-level logprobs during generation, while Alemohammad et al. [2024] show that biased or top-K sampling – retaining only the highest-scoring synthetic samples each generation – can slow collapse when a quality score derived from model internals is available. All three methods require access to the generating model's output distribution.

(B) External-oracle methods. Feng et al. [2025] argue that scaling up with synthesised data requires an external verifier – a stronger model or human annotator – to score and filter synthetic samples. Yi et al. [2025] extend this line with convergence guarantees for verifier-based filtering, but identify a critical caveat: long-run convergence is toward the *verifier's* biased knowledge centre, not the true data distribution, unless the verifier is perfect. This finding motivates the search for estimators that introduce no external model bias.

(C) Real-data accumulation. Gerstgrasser et al. [2024] prove that accumulating all prior real and synthetic data yields a finite error bound and validate this empirically. Bertrand et al. [2024] show that retaining a fixed fraction of real data each generation guarantees stability, and Fu et al. [2025] establish that even a constant proportion of real data ensures convergence in transformers. All three approaches require continued access to human-authored data at each training round.

Each type of mitigation thus requires either model internals (A), a stronger external model (B), or curated human data (C), imposing an extra burden to mitigate model collapse regardless of the approach taken. In this work we introduce a novel filter in category (D): $\hat{H}_K$ is computed from raw

[1] `https://github.com/lewismath/no-model-required`

text alone – no model, no API, no GPU, no reference corpus – making it, to our knowledge, the only completely model-free and reference-free model collapse filter proposed to date.

Relationship to data deduplication. Min-hash near-deduplication [Lee et al., 2022, Penedo et al., 2024] is the standard model-free text-quality filter in large-scale pretraining pipelines, and is the one existing method that, like $\hat{H}_K$-filtering, requires neither model access nor an external verifier. It removes near-duplicate *documents* from the training corpus (a between-document operation), whereas $\hat{H}_K$-filtering selects documents by within-document phrase variety (a within-document operation). The two approaches are complementary: deduplication prevents the corpus from containing many copies of the same text, while $\hat{H}_K$-filtering ensures each document has not collapsed into internally repetitive structure. A combined pipeline (deduplicate first, then $\hat{H}_K$-filter) would address both failure modes.

The Kontoyiannis estimator. Kontoyiannis et al. [1998] prove that the estimator

$$\hat{H}_K = \frac{n \log_2 n}{\sum_{i=1}^{n} \Lambda_i} \tag{1}$$

converges almost surely to the entropy rate H of any stationary ergodic process – no parametric assumptions on the token distribution are required – where Λ_i is the length of the shortest prefix of $x_i, \ldots, x_n$ that does not appear as a contiguous substring in $x_1, \ldots, x_{i-1}$; near the end of a document, where no such non-appearing prefix may exist, Λ_i is bounded by the remaining sequence length $n - i$, so the sum in Equation (1) remains well-defined without special-casing. Reported $\hat{H}_K$ values use the `ProcessEntropy` package's tokenizer[2]: an NLTK TweetTokenizer pass, lower-cased, with non-alphanumeric characters stripped from within each token [South et al., 2022]. Real-time filtering decisions during fine-tuning (Section 5) instead use a faster whitespace tokenization (lower-case; non-alphanumeric characters replaced with whitespace) for computational efficiency; the two agree closely in practice (Spearman $\rho = 0.997$ on a held-out sample), so the choice does not affect which documents are selected. The estimator's consistency is grounded in the Shannon–McMillan–Breiman theorem [Cover and Thomas, 2006]: for a stationary ergodic source, $-\frac{1}{n} \log_2 P(X_1, \ldots, X_n) \to H$ almost surely, and Λ_i serves as a nonparametric proxy for local self-information at position i – long matches signal low surprise; short matches signal high surprise. The match lengths also connect $\hat{H}_K$ to Lempel–Ziv (LZ) data compression [Ziv and Lempel, 1977, 1978]: each Λ_i identifies the shortest phrase at position i not resolvable from prior context – the elementary step of LZ parsing – so $\hat{H}_K$ is interpretable as a normalised LZ compression rate. While compression-based statistics have been used for general data quality assessment, to our knowledge no prior work has applied $\hat{H}_K$ or any LZ-derived metric as a model collapse mitigation filter. At finite n, $\hat{H}_K$ is upward-biased: early positions have little past context, so Λ_i is artificially small and the denominator is deflated. We control for this by computing $\hat{H}_K$ on the first 1,500 word tokens of each document and discarding shorter texts (see Section 3); the residual finite-sample bias is reported as a limitation in Section 6. The Kontoyiannis estimator provides a model-free measure of sequential complexity that depends only on the text itself [Bagrow et al., 2019, Smart et al., 2024].

Fine-tuning. We use QLoRA [Dettmers et al., 2023] – 4-bit NF4 quantisation with low-rank adaptation [Hu et al., 2022] – for all fine-tuning experiments. QLoRA makes multi-generation fine-tuning tractable on a single A100 GPU, enabling the iterative collapse and filtering experiments described in Sections 4 and 5.

3 $\hat{H}_K$ as a Proxy for Logprob Entropy

We first ask whether $\hat{H}_K$ is a suitable proxy for H_L, by comparing the two metrics on a set of synthetic documents.

3.1 Cross-domain and cross-temperature validation

We generated 200 documents using GPT-4o across 4 domains (encyclopedic, creative, scientific, conversational), 2 temperatures ($T \in \{0.5, 1.0\}$; higher temperatures were excluded in preliminary

[2] `https://github.com/tobinsouth/ProcessEntropy`

testing because too large a fraction of sampled tokens fell outside the top-20 returned logprobs), and 25 topics per domain; 197 documents passed a reliability filter requiring that fewer than 5% of a document's tokens have unrecoverable (invalid) logprob coverage. For each document we computed $\hat{H}_K$ via Equation (1) applied to the first 1,500 word tokens, and H_L, the mean per-token Shannon entropy of the generator's next-token distribution, renormalised over the top-20 returned logprobs.

An OLS regression of $\hat{H}_K$ on H_L with domain fixed effects yields $\beta = 0.924$ ($R^2 = 0.746$, $p < 0.0001$; Figure 2a shows the $T = 1.0$ subset). A likelihood-ratio test for domain-specific slopes fails to reject equality ($p = 0.41$), confirming a *domain-invariant slope*: the relationship between $\hat{H}_K$ and H_L is consistent across domains up to a domain-level intercept offset. Domains differ in baseline entropy – creative and conversational text is more varied; scientific text more constrained – but $\hat{H}_K$ scales with H_L at the same rate in each.

The domain-invariant slope has a practical consequence for data filtering. Because filtering requires only the *rank order* of documents by entropy (selecting the top-k), not absolute calibration, a single $\hat{H}_K$ computation suffices for within-domain ranking. In a mixed-domain corpus, domain stratification remains necessary to prevent over-selection from naturally higher-entropy domains.

3.2 Cross-model validation

The cross-domain experiments above use the same model (GPT-4o) for both generation and H_L scoring. Here we test whether the $\hat{H}_K$/H_L relationship holds when H_L is computed by a different model family altogether. We generated additional corpora using Gemini 1.5 Pro [Gemini Team, Google, 2024] and scored all documents – both GPT-4o-generated and Gemini-generated – using Qwen2.5-3B [Yang et al., 2024] as an independent scorer. Both the GPT-4o/Qwen and Gemini/Qwen pairings reproduce the positive $\hat{H}_K$/H_L relationship (Figure 2b), confirming that $\hat{H}_K$ measures a property of the text itself, not of any particular model's distribution over that text.

3.3 Scope: proxy validity on collapse-experiment data

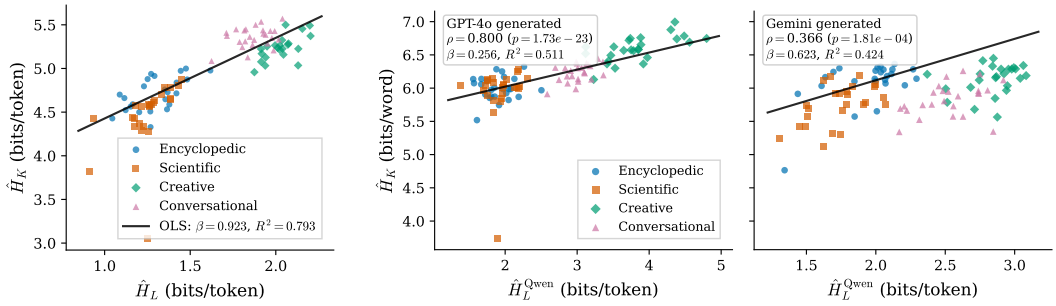
The validations above use text generated under controlled conditions (varied temperature, well-distributed topics). To assess whether the relationship extends to the data used in the collapse experiment itself, we compute $\hat{H}_K$ and H_L on the 80 generation-0 documents produced by Llama-3.1-8B-Instruct (the base model, before any fine-tuning). The correlation is near zero and non-significant ($r = 0.18$, $p = 0.11$, $n = 80$). This is not a failure of the estimator but a known property of the base model's output distribution: as documented in Figure 8, the base Llama model produces a wide bimodal H_L distribution (SD = 2.17) driven by topic-dependent familiarity rather than document-level diversity. Within each topic, H_L is nearly constant; between topics it varies by up to 7 bits. $\hat{H}_K$, which measures within-document phrase variety, is orthogonal to this between-topic variation. The proxy relationship holds when H_L reflects genuine document-level entropy variation, but not when it is dominated by topic-familiarity confounds – and this is the root cause of H_L-filtering's failure in Section 5.

4 Detecting Model Collapse Without Logprobs

Having established $\hat{H}_K$ as a cross-sectional proxy for H_L, we next ask whether it tracks collapse *over generations*. Under iterative rephrasing, $\hat{H}_K$ and H_L co-decline at indistinguishable rates ($\delta = -0.004$, $p = 0.59$), but within-generation rank concordance is near zero (Spearman $\rho \approx 0$), as rephrasing homogenises surface statistics across documents and eliminates the variance needed to discriminate between estimators. The fine-tuning regime is more informative, as we explore below.

4.1 Iterative fine-tuning collapse

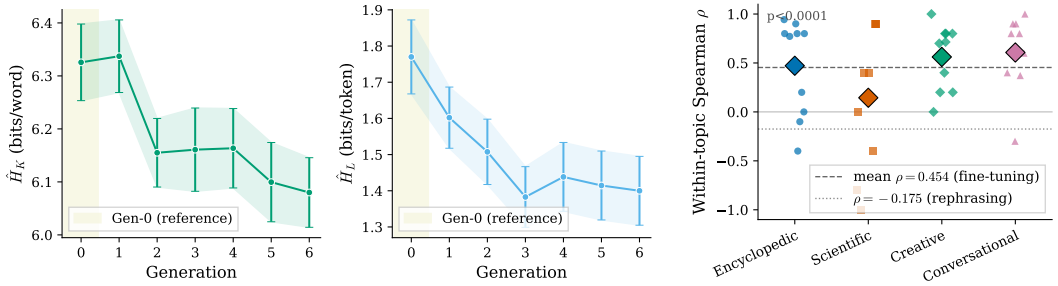
Fine-tuning collapse operates through a different mechanism than rephrasing. Rather than paraphrasing existing content, the model's *generative style* drifts: text exhibits increased phrase repetition, reduced vocabulary diversity, and narrower syntactic range (see examples in Appendix A). We conducted a 6-generation QLoRA fine-tuning experiment (Llama-3.1-8B-Instruct, unfiltered training



(a) Cross-domain validation: GPT-4o corpus ($T = 1.0$, $n = 97$).

(b) Cross-model validation: GPT-4o and Gemini corpora, both scored by Qwen2.5-3B.

Figure 2: $\hat{H}_K$ versus H_L across domains and model pairs. Points are coloured by domain. The slope is domain-invariant in both panels: $\hat{H}_K$ scales with H_L at a consistent rate regardless of domain, generating model, or scoring model.



(a) Iterative rephrasing collapse: mean ± SEM of $\hat{H}_K$ and H_L across rephrasing generations. Both estimators decline together ($\delta = -0.004$, $p = 0.59$). Shaded region: gen-0 reference (length-mismatched).

(b) Iterative fine-tuning collapse: within-topic Spearman ρ under fine-tuning collapse. Each point is one topic; diamonds show domain means.

Figure 3: $\hat{H}_K$ as a collapse detector under two regimes. Under rephrasing (a), $\hat{H}_K$ and H_L co-decline but within-generation concordance is ≈ 0. Under fine-tuning (b), within-topic concordance is strongly positive ($\rho = +0.454$, $p < 0.0001$), driven by phrase-level repetition that $\hat{H}_K$ directly measures.

data) and computed within-topic $\hat{H}_K/H_L$ concordance across generations for each of 39 topics (one topic was excluded due to insufficient within-generation variance for reliable rank correlation).

Under fine-tuning collapse, within-topic Spearman concordance is $\rho = +0.454$ ($p < 0.0001$), with 30 of 39 topics showing positive concordance (Figure 3b). This contrasts sharply with the rephrasing scenario ($\rho = -0.175$, non-significant). The explanation is that fine-tuning collapse changes the surface statistics of generated text in ways $\hat{H}_K$ directly measures – repetitive phrase structure, reduced n-gram diversity – whereas rephrasing preserves the within-generation homogeneity that masks both estimators. This result establishes $\hat{H}_K$ as a valid *detector* of fine-tuning collapse specifically, and motivates its use as a filter in that regime.

5 $\hat{H}_K$ as a Training Data Filter

Having established that $\hat{H}_K$ is a good collapse detector, we now consider our central question: can $\hat{H}_K$ be used to filter training data and mitigate collapse; and, if so, does it outperform H_L, the most established model-access-requiring baseline?

Table 1: Generation-6 outcomes under three training-data conditions. Diversity metrics show % change relative to unfiltered. p-values from Welch t-tests. $n = 80$ documents per condition.

Metric	Unfilt.	H_L-filt.	$\hat{H}_K$-filt.	$p\,(\hat{H}_K$ v. U)	$p\,(H_L$ v. U)	$p\,(\hat{H}_K$ v. $H_L)$
$\hat{H}_K$ (bits/word)	0.79	1.16	**1.43**	<0.001	0.008	<0.001
Distinct-3 (%)	–	$+7\,\%$	**+42 %**	<0.001	0.23	<0.001
Vocabulary (%)	–	$+4\,\%$	**+30 %**	<0.001	0.31	<0.001
Rep-4 (%)	–	$-3\,\%$	**-19 %**	<0.001	0.27	<0.001

5.1 Experimental design

We run three parallel 6-generation QLoRA fine-tuning chains on Llama-3.1-8B-Instruct. At each generation, 80 documents are generated across 4 domains (10 topics $\times$ 2 documents per topic per domain); 40 are selected for training according to the condition:

Unfiltered**:** random selection (seed fixed per generation);

H_L-filter**:** top 40 by mean per-token Shannon entropy (logprob-based surplexity; see Section 1 for the distinction between H_L and surplexity proper); and

$\hat{H}_K$-filter**:** top 40 by Kontoyiannis entropy rate (text-only).

Selection is domain-stratified (10 documents per domain), so both filters operate within domain. Generation-0 documents are drawn from the base model and are shared across conditions. Training uses 4-bit NF4 QLoRA (rank 16, $\alpha = 16$, learning rate 2×10^{-4}, cosine schedule, 3 epochs per generation, effective batch size 8). Documents selected for training are truncated at the onset of 4-gram repetition; this preprocessing does not apply to generated outputs, which are evaluated as-is (see Appendix A). The experiment produces 1,520 documents in total: 80 shared generation-0 documents plus 3 conditions $\times$ 6 further generations $\times$ 80 documents.

Reproducibility. Generation uses a fixed decoding temperature of 1.0 throughout. Prompts follow a fixed template with a domain-specific system prompt (e.g., for encyclopedic: "You are writing an encyclopaedia. Write in a neutral, factual, Wikipedia-style tone.") and a shared user prompt, "Write a detailed, long-form piece about the following topic. Write at least 3,000 words. Do not stop early. Topic: {topic}." The topic set (10 per domain $\times$ 4 domains) is identical across all six generations and across all three conditions: the only thing that varies by condition is which 40 of the 80 candidate documents are selected for training at each generation and, from generation 1 onward, which fine-tuned checkpoint generates the following round's candidate documents.

5.2 Results

Table 1 summarises outcomes at generation 6 on four metrics. $\hat{H}_K$ (bits/word) and three text-diversity measures – Distinct-3 [Li et al., 2016], vocabulary size, and intra-document 4-gram repetition rate (Rep-4) – are computed on the 80 held-out generation-6 documents per condition. Diversity metrics are shown as percentage change relative to the unfiltered condition.

$\hat{H}_K$-filtering slows collapse. At generation 6, $\hat{H}_K$-filtering retains 1.43 bits/word versus 0.79 for unfiltered ($p < 0.001$) and 1.16 for H_L-filtering ($p < 0.001$). On all three text-diversity metrics, $\hat{H}_K$-filtering produces large, significant improvements over unfiltered training: $+42\%$ more unique trigrams, $+30\%$ larger vocabulary, and -19% less intra-document repetition (all $p < 0.001$). These gains are consistent across all four domains (domain-level trajectories and full metric suites across generations are shown in Appendix B). Illustrative generation-6 documents from the unfiltered and $\hat{H}_K$-filter conditions are shown in Appendix A.

H_L-filtering provides no text-diversity benefit. Despite having access to the model's log-probabilities, H_L-filtering produces no significant improvement on any of the three text-diversity metrics ($p > 0.23$ on all). It does outperform unfiltered training on the $\hat{H}_K$ metric itself ($p = 0.008$), but the effect is less than half that of the $\hat{H}_K$-filter. The failure has a clear mechanistic explanation:

after even one fine-tuning step, the H_L distribution compresses dramatically (gen-0 SD $= 2.17$; gen-1 SD $= 0.13$), collapsing the ranking signal the filter depends on. Within a generation whose outputs are already near-uniform in H_L, the filter degenerates to near-random selection.

Mixed-effects analysis. A mixed-effects regression ($\hat{H}_K \sim$ generation $\times$ condition $+ (1|\text{topic})$, baseline: unfiltered) decomposes the trajectory into a starting-quality component and a rate-of-collapse component. $\hat{H}_K$-filtering produces a significant intercept shift ($\beta = +0.37$ bits, $p = 0.001$); because all conditions share the same generation-0 documents, this reflects the shape of the fitted trajectories rather than a pre-existing quality difference. The slope improvement is positive but does not reach significance (Δslope $= +0.058$ bits/gen, $p = 0.067$). H_L-filtering produces neither a significant intercept shift nor a slope improvement ($p > 0.1$ on both).

The filters select different documents. Within-domain Jaccard similarity between the H_L-filter and $\hat{H}_K$-filter selection sets has a mean of 0.35 (range 0.18–0.54 across generations and domains). The two criteria do not approximate each other: $\hat{H}_K$ selects on phrase-level novelty, while H_L selects on token-level model uncertainty. The success of $\hat{H}_K$-filtering despite this low overlap suggests that phrase-level repetition, not token-level surprise, is the more informative axis for collapse mitigation – consistent with the observation that fine-tuning collapse manifests primarily in surface repetition.

$\hat{H}_K$ outperforms simpler text-only baselines. A natural objection is that any per-document diversity statistic might replicate $\hat{H}_K$'s advantage – for instance, type–token ratio (TTR) or bigram distinctiveness (Distinct-2). We assess this by simulating both filters on the generation-0 pool, where all conditions share the same 80 documents and the comparison is exact. TTR@500-filtering selects training documents with mean $\hat{H}_K = 3.37$; Distinct-2-filtering selects mean $\hat{H}_K = 3.32$; $\hat{H}_K$-filtering selects mean $\hat{H}_K = 3.98$ – a 18–20% advantage. Despite moderate positive correlations between TTR and $\hat{H}_K$ in the pool (Pearson $r = 0.54$), the selection sets overlap only 43% by Jaccard, indicating that $\hat{H}_K$ recovers phrase-level structure that per-document surface statistics do not capture. The difference is structural: TTR and Distinct-2 aggregate over short local windows, whereas $\hat{H}_K$ is sensitive to long-range match structure – precisely the property that fine-tuning collapse erodes. We repeated this static-pool comparison against two further model-free criteria: an off-the-shelf gzip-compression ratio (least-compressible 40 documents selected per domain) and an explicit 4-gram repetition-rate filter (lowest repeated-4-gram rate). Both correlate with $\hat{H}_K$ only moderately (gzip $r = +0.44$, $p < 0.001$; repetition rate $r = -0.36$, $p = 0.001$) – comparable to the TTR correlation above – and select substantially different document sets (Jaccard 0.36–0.38 against the $\hat{H}_K$-filter selection). As with the TTR/Distinct-2 comparison, this establishes that $\hat{H}_K$ is related to but not redundant with either criterion on the static pool, but not downstream collapse-resistance, which we test directly by running both as genuine six-generation training filters, replacing $\hat{H}_K$ in the filtering criterion but otherwise identical to Section 5. At generation 6, $\hat{H}_K$-filtering significantly outperforms both on every metric ($\hat{H}_K$, Distinct-3, Rep-4, vocabulary; all $p < 0.005$, most $p < 10^{-4}$). The two static-pool-comparable criteria do not perform alike as real filters, however: gzip-filtering gives a modest but genuine diversity benefit over unfiltered training (Distinct-3 $+17\%$, vocabulary $+17\%$, Rep-4 -7%; all $p < 0.03$), while repetition-rate filtering gives none (all four metrics indistinguishable from unfiltered, $p > 0.24$) despite directly targeting repetition. $\hat{H}_K$'s advantage over simpler model-free heuristics is therefore not just a static-pool artefact: it persists, and in repetition-rate filtering's case is entirely absent for the alternative, under the same six-generation training loop that constitutes this paper's central result.

External validation. We scored gen-6 documents from all conditions using GPT-2 [Radford et al., 2019], a 124M-parameter model with no relation to training or generation pipelines. GPT-2 entropy shows the same directional ordering ($\hat{H}_K$-filtered $> H_L$-filtered $>$ unfiltered) but does not reach significance ($p = 0.11$), limiting its value as independent validation; GPT-2's discriminative power on long-form text is insufficient to detect the effect at the individual-document level. We also evaluated each generation's fine-tuned checkpoint on HellaSwag [Zellers et al., 2019] using Llama-3.1-8B-Instruct (Figure 4b). Accuracy drops sharply from generation 0 to generation 1 ($0.729 \rightarrow\approx 0.672$) and then plateaus near 0.652–0.657 with no significant difference between conditions at any generation. Thus $\hat{H}_K$-filtering preserves text diversity but does not protect general

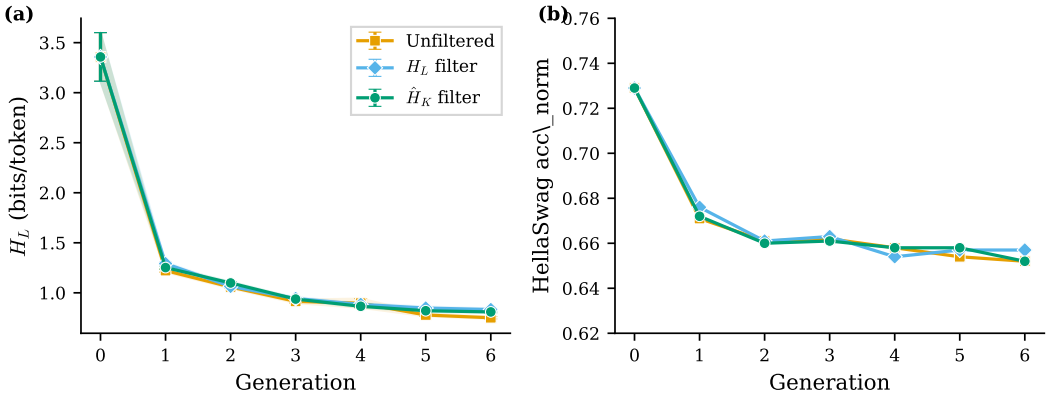


Figure 4: **(a)** H_L (logprob entropy) trajectories across six fine-tuning generations for the three filtering conditions (± 1 SEM bands). All collapse to low H_L by gen-1 and track nearly identically, illustrating why the H_L filter fails: post-collapse variance is too small to discriminate. **(b)** HellaSwag normalised accuracy (Llama-3.1-8B-Instruct) across six generations. All three conditions degrade from 0.729 (gen-0) to ≈ 0.652–0.657 (generation 6) with no significant difference between conditions, indicating that text-diversity preservation does not translate to preserved general reasoning ability.

reasoning ability: the capability degradation introduced by iterative fine-tuning is uniform across all three conditions. Two further downstream benchmarks corroborate this: 5-shot MMLU and 8-shot GSM8K (1,000-example subsamples each) on the same checkpoints show no significant difference between conditions at generation 6 (MMLU: 67.6–67.6%; GSM8K: 70.7–71.2%; all pairwise χ^2 tests $p > 0.8$), nor a significant decline from generation 0 to generation 6 within any single condition ($p > 0.15$) – consistent with HellaSwag, general capability retention is unaffected by which text-diversity filter is used, at this scale and over this many generations.

6 Discussion

Why $\hat{H}_K$-filtering outperforms H_L-filtering. The Kontoyiannis estimator directly penalises repeated phrase structure: a document with many recurring n-grams exhibits long match lengths and hence a low $\hat{H}_K$. Fine-tuning collapse manifests primarily as increased phrase repetition, so $\hat{H}_K$-filtering targets collapse's dominant surface signature directly. H_L, by contrast, measures token-level uncertainty in the scorer model – a signal that is destroyed after a single fine-tuning step, as shown in Section 5. At generation-0, H_L primarily reflects topic familiarity – the wide bimodal distribution documented in Figure 8 – rather than document diversity, so a criterion uncorrelated with document quality at generation-0 cannot select high-quality training data regardless of downstream variance. The result suggests that collapse is fundamentally a phrase-repetition phenomenon, and that phrase-level statistics are better positioned to detect it than token-level uncertainty. A literal reimplementation of surplexity confirms this mechanistic prediction directly: its median absolute deviation across the shared generation-0 pool is 4.40, falling to 0.40–0.49 (all three conditions) after one fine-tuning step – a 9–11-fold narrowing, comparable to H_L's own collapse. (Generation-0 surplexity is heavy-tailed – a handful of unfamiliar-topic documents receive extreme values once surprise is exponentiated into perplexity – so median and MAD, not mean and SD, are the meaningful summary here.)

Filter design determines direction of effect. Not all filters slow collapse: Guo et al. [2024] find that linguistic acceptability filtering (a RoBERTa classifier trained on CoLA) *accelerates* diversity decline, since such filters retain conventional, high-frequency text and discard the inventive samples most worth keeping. $\hat{H}_K$-filtering operates on the opposite principle, selecting for the phrase-level novelty that collapse erodes.

The Jaccard gap and Simpson's paradox. The reason why the two filters select such different documents (Jaccard ≈ 0.35) when $\hat{H}_K$ and H_L are well-correlated at the corpus level ($R^2 = 0.746$) is

a form of Simpson's paradox. The *between*-domain correlation is strong: creative and conversational text have both high H_L and high $\hat{H}_K$; scientific text has both low. But *within* a domain and generation – which is where filtering operates, due to domain stratification – the residual correlation is near zero. The estimators measure genuinely different properties of intra-domain text variation; $\hat{H}_K$'s phrase-level novelty proves more informative for collapse resistance.

Practical implications. $\hat{H}_K$ filtering requires no model access, no API calls, and no GPU: scoring a 1,500-word document takes under 10 ms, and a pipeline producing 10^5 candidate documents per round incurs negligible, trivially parallelisable overhead. This makes collapse-resistant data curation practical where logprob access is unavailable: closed-source API distillation, cross-organisation data sharing, and federated training.

Limitations. Most importantly, $\hat{H}_K$-filtering preserves text diversity but does not protect general reasoning ability: HellaSwag accuracy degrades equally across all three filtering conditions (Figure 4b), and MMLU and GSM8K show the same pattern, so our claim is scoped to distributional fidelity of generated text, not downstream task performance. The filtering advantage is also specific to fine-tuning collapse: under iterative rephrasing, $\hat{H}_K$ tracks the aggregate trajectory but cannot distinguish among documents within a generation, and is therefore unsuitable as a within-generation filter in that setting. The diversity gains measured here are lexical: MAUVE scores [Pillutla et al., 2021] (GPT-2-Large features, first 512 tokens) are near-ceiling and uniform across conditions at generation 6 (0.987–0.993), indicating embedding-space similarity is preserved and $\hat{H}_K$'s advantage is specific to within-document match-length structure. A separate concern is quality: selecting for match-length novelty could favour incoherent or off-topic text over genuinely diverse text. We address this directly with an LLM-judge pass (GPT-4o-mini, blind to condition) scoring all 240 generation-6 documents (80 per condition) on two 1–5 scales, coherence and instruction-following, reading the full document so that late-document degeneration is not missed. $\hat{H}_K$-filtering scores highest of the three conditions on both: coherence 2.35 versus 2.20 for unfiltered ($p = 0.058$) and 2.29 for H_L-filtering; instruction-following 3.23 versus 3.00 for unfiltered ($p = 0.020$) and 3.15 for H_L-filtering. If anything, then, $\hat{H}_K$-filtering's diversity advantage comes with a small quality benefit rather than a coherence cost. Finally, our collapse experiment uses a single model family (Llama-3.1, 8B); replicating with Llama-3.2-3B-Instruct (same corpus and protocol), $\hat{H}_K$-filtering again retains more diversity at generation 6 (2.91 bits/word versus 2.12 unfiltered, $p < 0.001$, and 2.41 for H_L-filtering, $p = 0.004$; H_L-filtering again non-significant on all text-diversity metrics), with the $\hat{H}_K$-vs-H_L comparison itself reaching significance on the $\hat{H}_K$ metric ($p = 0.004$ vs $p = 0.088$ at 8B), providing stronger evidence the two filters are genuinely different. Beyond replication, the Kontoyiannis cross-entropy rate $\hat{h}_\times(T \mid S)$ is a natural complement to $\hat{H}_K$, selecting for documents that have diverged from prior-generation content; the same signal could operationalise – as an automatic, text-only proxy – the blind-writing independence phase Chen et al. [2026] identify as the most effective structural intervention against diversity collapse.

Broader impacts. The primary benefit is making collapse-resistant data curation accessible without model access, reducing a practical barrier for organisations running iterative fine-tuning under closed-source or resource-constrained conditions, and counteracting homogenisation of AI-generated web content. The main negative consequence is indirect: making fully-synthetic training pipelines more viable may reduce the incentive to collect or compensate human-authored training data; practitioners deploying such pipelines at scale should weigh this trade-off.

7 Conclusion

We have shown that the Kontoyiannis entropy rate estimator $\hat{H}_K$ – a model-free, match-length measure of sequential complexity – is an effective training-data filter against fine-tuning-driven model collapse: at generation 6 it produces 42% more unique trigrams, 30% larger vocabulary, and 19% less intra-document repetition than unfiltered training, while logprob-based filtering shows no significant diversity benefit. For iterative pipelines where model internals are inaccessible, $\hat{H}_K$-filtering offers a strong, zero-cost baseline.

Acknowledgments and Disclosure of Funding

This research was supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP210103700). Computational resources provided by the Phoenix HPC cluster at Adelaide University.

References

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. In *International Conference on Learning Representations*, 2024.

James P. Bagrow and Lewis Mitchell. The quoter model: a paradigmatic model of the social flow of written information. *Chaos*, 28:075304, 2018.

James P. Bagrow, Xipei Liu, and Lewis Mitchell. Information flow reveals prediction limits in online social activity. *Nature Human Behaviour*, 3:122–128, 2019.

Quentin Bertrand, Avishek Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. In *International Conference on Learning Representations*, 2024.

Martin Briesch, Dominik Sobania, and Franz Rothlauf. Large language models suffer from their own output: An analysis of the self-consuming training loop. *arXiv preprint arXiv:2311.16822*, 2023.

Nuo Chen, Yicheng Tong, Yuzhe Yang, Yufei He, Xueyi Zhang, Qingyun Zou, Qian Wang, and Bingsheng He. Diversity collapse in multi-agent LLM systems: Structural coupling and collective failure in open-ended idea generation. *arXiv preprint arXiv:2604.18005*, 2026.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Abhimanyu Dubey et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yunzhen Feng, Elvis Dohmatob, Pu Yang, François Charton, and Julia Kempe. Beyond model collapse: Scaling up with synthesized data requires verification. In *International Conference on Learning Representations*, 2025.

Shi Fu, Yingjie Wang, Yuzhu Chen, Xinmei Tian, and Dacheng Tao. A theoretical perspective: How to prevent model collapse in self-consuming training loops. In *International Conference on Learning Representations*, 2025.

Daniele Gambetta, Gizem Gezici, Fosca Giannotti, Dino Pedreschi, Alistair Knott, and Luca Pappalardo. Learning by surprise: Adaptive mitigation of model collapse in large language models. *ACM Transactions on Intelligent Systems and Technology*, 2026. doi: 10.1145/3828663.

Gemini Team, Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David Donoho, and Sanmi Koyejo. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*, 2024.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Ioannis Kontoyiannis, Paul H. Algoet, Yurii M. Suhov, and Aaron J. Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Transactions on Information Theory*, 44(3):1319–1327, 1998.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8424–8445, 2022.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, 2016.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, 2002.

Guilherme Penedo et al. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828, 2021.

Tyson Pond, Saranzaya Magsarjav, Tobin South, Lewis Mitchell, and James P Bagrow. Complex contagion features without social reinforcement in a model of social information flow. *Entropy*, 22 (3):265, 2020.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

Bridget Smart, Matthew Roughan, and Lewis Mitchell. The entropy rate of linear additive markov processes. *Plos one*, 19(4):e0295074, 2024.

Tobin South, Bridget Smart, Matthew Roughan, and Lewis Mitchell. Information flow estimation: a study of news on Twitter. *Online Social Networks and Media*, 31:100231, 2022.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of LREC*, 2020.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Bingji Yi, Qiyuan Liu, Yuwei Cheng, and Haifeng Xu. Escaping model collapse via synthetic data verification: Near-term improvements and long-term convergence. *arXiv preprint arXiv:2510.16657*, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Xuekai Zhu et al. How to synthesize text data without model collapse? In *International Conference on Machine Learning*, 2025.

Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977.

Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.

Appendices

A Qualitative Examples of Fine-Tuning Collapse

The following excerpts illustrate the surface-level deterioration that $\hat{H}_K$ measures. Each example shows the same document topic at generation 0 (shared across all conditions) and generation 6 under the unfiltered and $\hat{H}_K$-filter conditions. $\hat{H}_K$ values are computed on the first 1,500 word tokens per document.

Example 1 (Conversational): ***Experiences with unexpected kindness from strangers***

Generation 0 ($\hat{H}_K = 4.72$ bits/word)

A Surprising Revelation: Unexpected Kindness from Strangers
As I sit here, reflecting on the countless moments that have shaped my life, I find myself pondering a peculiar yet profound truth. It's a truth that often goes unspoken, yet resounds loudly in my mind: unexpected kindness from strangers can be the spark that ignites a chain reaction of compassion, empathy, and connection. In this essay, I'll explore the unexpected kindness that I've experienced firsthand, and I'll invite you to join me on a journey of self-discovery, as we explore the transformative power of kindness from strangers.
The Power of Random Acts of Kindness
As I was walking home from work one day, lost in thought, I stumbled upon a small café. I had been there before, but today, I spotted a poster advertising a free lunch for low-income families. My heart ached as I watched a group of homeless individuals, struggling to find any form of sustenance, pushing carts filled with fresh produce and leftovers. [. . .] As I took a bite, I watched a young boy, no more than 10 years old, sitting across from me, staring intently at a drawing on the table. He seemed to be struggling to put the pencil to paper, his eyes welling up with tears.

Generation 6, unfiltered ($\hat{H}_K = 0.08$ bits/word)

The Power of Unexpected Kindness: A Reflection on the Beauty of Human Connection
As I sit here, reflecting on the past year, I find myself pondering the many unexpected moments of kindness that have touched my life. It's a theme that has become increasingly important to me, as I've come to realize the significance of the small acts of kindness that can have a profound impact on our lives. In this essay, I'll share with you some of the most memorable and heartwarming experiences I've had with unexpected kindness from strangers. [. . .]
[. . . after eight early sections, the document collapses into verbatim-repeated paragraphs:]
The Unexpected Support System
One of the most surprising experiences I've had with unexpected kindness was when I received a message from a stranger, who asked me to join them for a support group. They invited me to join a group of people who were going through similar struggles, and they promised that they would be there to support me. I was hesitant at first, but I decided to take a chance. I attended the support group, and it was a turning point for me. I felt a sense of connection with the people around me, and I felt like I was not alone. [. . .] I realized that sometimes, the most unexpected gifts can come from the most unexpected places.
The Unexpected Gift of Empathy
One of the most surprising experiences I've had with unexpected kindness was when I received a message from a stranger, who asked me to join them for a self-discovery retreat. They invited me to join a group of people who were going through similar struggles, and they promised that they would be there to support me. I was hesitant at first, but I decided to take a chance. I attended the retreat, and it was a turning point for me. [. . .]
[The above two paragraphs alternate and repeat verbatim approximately eight further times until the document length limit is reached.]

Generation 6, $\hat{H}_K$-filter ($\hat{H}_K = 5.07$ bits/word)

The Power of Unexpected Kindness: Lessons from Strangers
I remember the day I stumbled upon a stranger who changed my life forever. It was a chilly winter morning, and I was walking home from work, lost in thought. The cold air hit me like a slap in the face, and I felt a surge of frustration and anger. [. . .]

As I turned a corner, I noticed a woman standing on the sidewalk, her eyes downcast and her hair disheveled. She was wearing a faded jacket and a pair of worn-out shoes, and she looked like she had been walking for hours. [...] She offered me the food, and I gratefully accepted, feeling a sense of gratitude towards this stranger.
The woman, who I later learned was a young mother on her way to a job interview, smiled at me as she handed me the food. "Thank you," she said, her voice barely above a whisper. "I'm so tired. I haven't had any breakfast in hours." I felt a surge of empathy for her, and I wanted to help her in any way I could.

Note that the $\hat{H}_K$-filtered example ($\hat{H}_K = 5.07$ bits/word) is a high-entropy outlier relative to the condition's generation-6 mean ($\hat{H}_K = 1.43$ bits/word); it is selected here to illustrate the qualitative contrast, not to represent the average filtered output.

The collapse in the unfiltered generation-6 document is qualitatively unmistakable: the second half consists of near-verbatim repeated paragraphs that contribute no new content. The $\hat{H}_K$ metric reflects this directly – long match lengths throughout the document's latter portion drive the estimate to near zero. The $\hat{H}_K$-filtered document opens with a specific, situated narrative, consistent with its $\hat{H}_K$ value exceeding generation 0 (the filter selected an unusually high-entropy document at generation 0 to begin the lineage, and the fine-tuned model preserved that stylistic register longer).

Example 2 (Scientific): ***Graphene oxide membranes for water desalination***

Generation 0 ($\hat{H}_K = 1.65$ bits/word)

Graphene Oxide Membranes for Water Desalination: A Review of their Mechanisms, Advantages, and Limitations
Water desalination remains a pressing issue worldwide, particularly in regions with limited access to freshwater resources. Traditional methods, such as reverse osmosis (RO), are energy-intensive and often costly. [...] Graphene oxide (GO) is a derivative of graphene, where oxygen atoms are replaced by hydroxide groups on the graphene surface, enhancing electrical conductivity, hydrolysis resistance, and water adsorption capacity.
Mechanisms of GO Membranes
The GO membrane is fabricated by depositing GO on a substrate, such as silicon dioxide (SiO_2) or carbon nitride, followed by deposition of a sacrificial layer such as indium tin oxide (ITO). The GO layer is then cured and thermally treated to form a composite membrane. The GO membrane exhibits unique properties: (1) *enhanced hydrophilicity* – surface energy significantly higher than pure graphene; (2) selective ion exclusion controlled by interlayer spacing [...]

Generation 6, unfiltered ($\hat{H}_K = 0.12$ bits/word)

Graphene Oxide Membranes for Water Desalination: A Review of the Current State of the Art and Future Directions
Graphene oxide membranes have emerged as a promising technology for desalination, particularly for removing salt and other minerals from seawater and brackish water. Graphene, a highly conductive and versatile two-dimensional material, has been extensively researched and developed for various applications, including energy, electronics, and water treatment. [...] Desalination is the process of removing salt and other minerals from seawater and brackish water to produce fresh water for human consumption, industry, and agriculture. [...] Graphene oxide membranes have been shown to be highly efficient in removing salt and other minerals from water, making them an attractive alternative to traditional desalination methods. [...]

Generation 6, $\hat{H}_K$-filter ($\hat{H}_K = 1.19$ bits/word)

Graphene Oxide Membranes for Water Desalination: A Review of the Current State of the Art and Future Directions
Graphene oxide membranes have emerged as a promising technology for water desalination, offering several advantages over traditional reverse osmosis (RO) and electrodialysis (ED) techniques. [...] The graphene oxide layer is typically 10–20 nanometres thick and consists of a mixture of graphene oxide, hydroxyl groups, and other functional groups. [...]

Permeation Properties of Graphene Oxide Membranes
Graphene oxide membranes have been shown to have exceptional permeation properties, allowing for the separation of ions and molecules with high selectivity. The permeation properties are influenced by the type of functional groups present, the thickness of the membrane, and the temperature and pressure conditions. [...]

This example illustrates a different collapse mode from Example 1. Rather than verbatim repetition, the unfiltered generation-6 document exhibits *semantic dilution*: the phrase "removing salt and other minerals from seawater and brackish water" appears three times in the opening 200 words, while specific mechanistic details present in generation 0 – fabrication steps, substrate materials, numbered membrane properties – are absent. The $\hat{H}_K$-filtered document partially recovers this specificity, retaining membrane thickness values (10–20 nm), functional group details, and permeation mechanism structure. Scientific-domain documents start with lower absolute $\hat{H}_K$ values than conversational ones (domain-level baseline effect, Section 3), so the absolute values are not directly comparable across examples; the relative contrast within each example is the relevant quantity.

B Supplementary Figures

This appendix provides additional figures supporting the main experimental results. The first two figures (Figures 5 and 6) expand the quantitative picture of Section 5 across multiple entropy estimators and text-diversity metrics. Figure 7 examines whether the filtering advantage is uniform across the four document domains or concentrated in particular content types. Figure 8 provides a mechanistic explanation for the failure of H_L-based filtering by characterising the generation-0 H_L distribution.

Three-estimator comparison (Figure 5). The main text reports results for $\hat{H}_K$ as the primary collapse metric and briefly notes the H_L trajectories. Figure 5 places all three estimators – $\hat{H}_K$, H_L^{Llama}, and the independent $H_L^{\text{GPT-2}}$ scorer – side-by-side across generations, making two patterns visible simultaneously. First, all three estimators agree on direction: the $\hat{H}_K$-filter condition retains higher entropy than H_L-filter, which in turn exceeds unfiltered, at generation 6. Second, the H_L^{Llama} panel shows the convergence that disables logprob filtering: after generation 0, the three condition means are nearly indistinguishable and their confidence bands overlap almost entirely, so a filter ranking documents by H_L within a generation operates on noise. The $\hat{H}_K$ panel exhibits a qualitatively different trajectory, with the $\hat{H}_K$-filter condition maintaining a visible separation from unfiltered throughout all six generations.

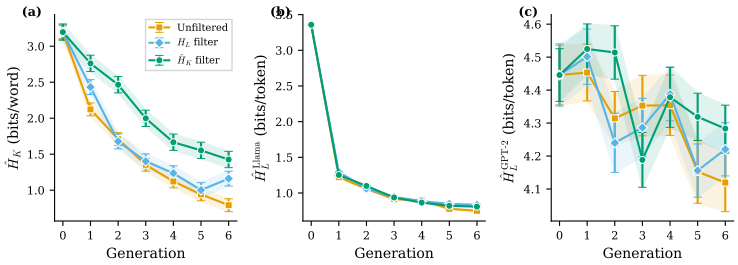


Figure 5: Three entropy metrics across six fine-tuning generations: $\hat{H}_K$ (text-only), H_L^{Llama} (Llama-3.1-8B scorer), and $H_L^{\text{GPT-2}}$ (independent GPT-2 scorer). All three metrics show the same directional ordering at generation 6. The H_L^{Llama} panel illustrates why logprob-based filtering fails: after generation 0, the three conditions are nearly indistinguishable, leaving the H_L-filter little variance to exploit.

Full text-diversity metric suite (Figure 6). Table 1 in the main text reports four summary metrics at generation 6. Figure 6 shows trajectories for six metrics across all generations, providing a more complete view of how diversity evolves over the course of the experiment. The within-document metrics (type–token ratio, Distinct-2, Distinct-3, vocabulary size, and Rep-4) all tell a consistent story:

the $\hat{H}_K$-filter condition (green) separates from the other two conditions by generation 1 and maintains a persistent advantage, while the H_L-filter condition (orange) tracks close to unfiltered on every within-document measure throughout. The Self-BLEU-4 [Papineni et al., 2002] panel (cross-document homogeneity) shows a more complex non-monotonic pattern in early generations, consistent with the discussion in Section 5: the $\hat{H}_K$-filter selects a statistically coherent cluster of documents, which briefly increases cross-document similarity. This does not fully resolve by generation 6: $\hat{H}_K$-filter's Self-BLEU-4 remains distinctly higher (0.086) than both unfiltered (0.075) and H_L-filter (0.060, both $p < 0.0001$, Welch's t-test) – a small, sustained cross-document homogeneity cost that partly offsets $\hat{H}_K$-filtering's much larger within-document diversity gains reported in Table 1.

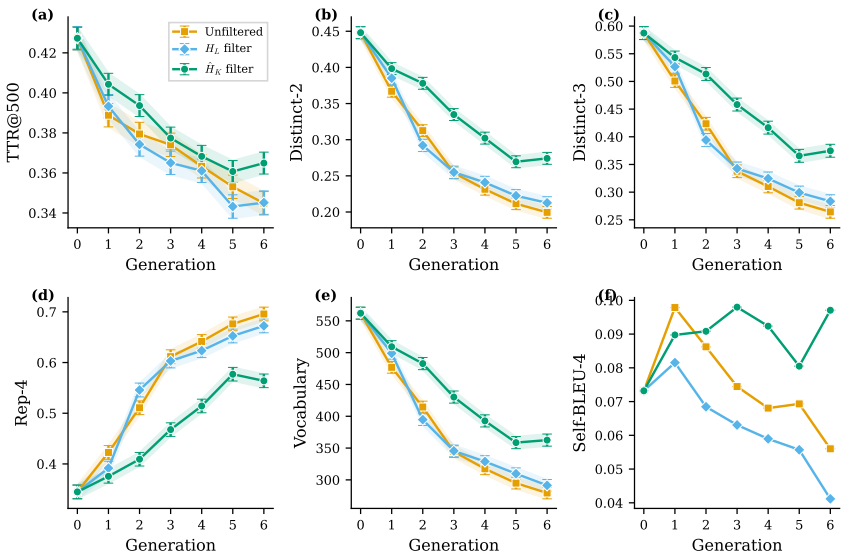


Figure 6: Six text-diversity metrics across generations under the three conditions. $\hat{H}_K$-filtering (green) maintains higher diversity on all within-document metrics (TTR, Distinct-n, vocabulary, Rep-4). Self-BLEU-4 (cross-document homogeneity) shows a non-monotonic pattern in early generations.

Domain-level breakdown (Figure 7). A natural concern is that the aggregate filtering advantage reported in Table 1 could be driven by one domain in which $\hat{H}_K$ happens to be a particularly good signal, while other domains show little or no benefit. Figure 7 plots $\hat{H}_K$ trajectories separately for each of the four domains – encyclopedic, scientific, conversational, and creative. The $\hat{H}_K$-filter advantage is present and directionally consistent across all four, though the magnitude varies: the gap is largest in the conversational domain (where collapse manifests most severely as phrase-level repetition, the exact phenomenon $\hat{H}_K$ measures) and smallest in the creative domain. No domain shows the H_L-filter outperforming unfiltered on $\hat{H}_K$, confirming that the null result for H_L-filtering is not an artefact of domain averaging.

Generation-0 H_L distribution (Figure 8). Section 5 argues that H_L-filtering fails because the within-generation H_L variance collapses after a single fine-tuning step, removing the ranking signal the filter depends on. Figure 8 documents the starting state that makes this explanation complete. At generation 0 – documents produced by the base Llama-3.1-8B-Instruct model before any fine-tuning – the H_L distribution is wide and bimodal (mean $= 3.36$, SD $= 2.17$). The bimodality reflects topic-dependent model familiarity: the base model assigns low entropy to topics well-represented in its pretraining data and high entropy to unfamiliar topics, producing two distinct modes rather than a smooth unimodal spread. After a single generation of fine-tuning on 40 documents, the distribution collapses to a narrow unimodal regime (SD ≈ 0.13, a 17-fold reduction), where all topics are now equally familiar to the fine-tuned model. From generation 1 onward, every document receives a similar H_L score, so the H_L-filter's ranking is determined by noise rather than signal. The wide generation-0 distribution also explains why the H_L-filter is uncorrelated with $\hat{H}_K$ at generation 0

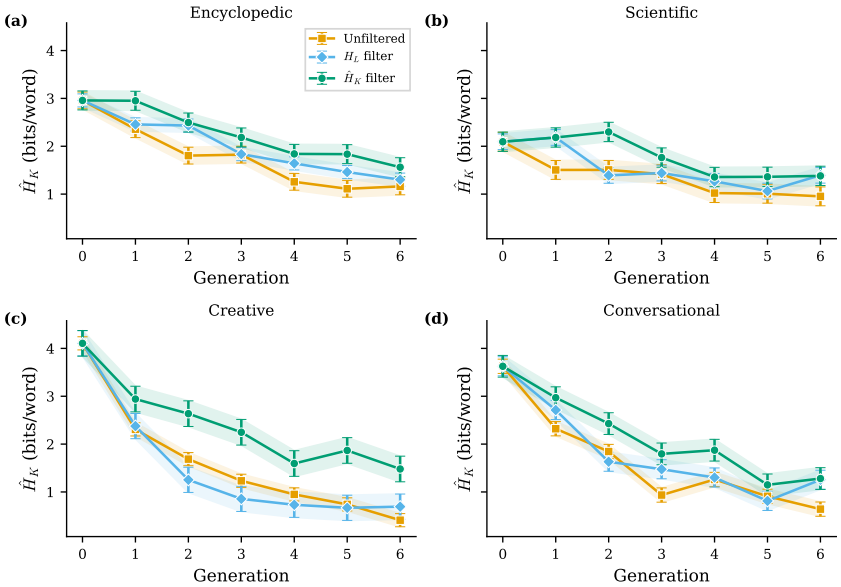


Figure 7: Domain-level $\hat{H}_K$ trajectories under the three conditions. The $\hat{H}_K$-filter advantage is consistent across all four domains.

($r = 0.18$, $p = 0.11$, Section 3): the bimodal H_L variation captures topic familiarity, not document diversity.

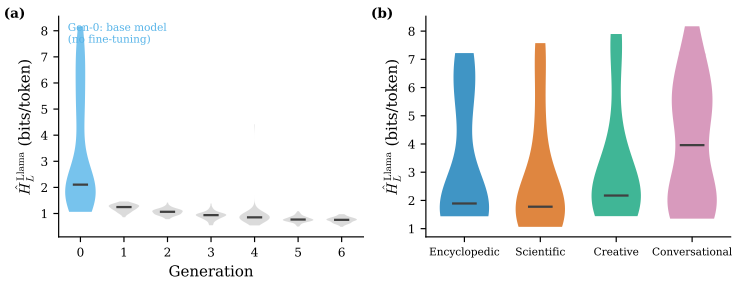


Figure 8: Generation-0 H_L distribution. The base Llama-3.1-8B-Instruct model exhibits a wide, bimodal H_L distribution (mean $= 3.36$, std $= 2.17$), driven by topic-dependent familiarity. After one fine-tuning step the distribution collapses to a narrow regime (std ≈ 0.13), eliminating the variance the H_L-filter relies on.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three claims – that $\hat{H}_K$ is a valid cross-domain entropy proxy, a fine-tuning collapse detector, and a superior training-data filter to H_L-based filtering – all of which are directly supported by the experimental results in Sections 3–5. Scope is appropriately scoped to the single-lineage, fully-synthetic fine-tuning setting; limitations on generality are stated in Section 6.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 contains a dedicated Limitations paragraph covering: (1) text-diversity preservation does not protect general reasoning ability (HellaSwag); (2) the filter is specific to fine-tuning collapse and fails under iterative rephrasing; (3) diversity gains are lexical (MAUVE scores are near-ceiling); (4) finite-sample upward bias of $\hat{H}_K$; (5) results are from a single model family (Llama-3.1) at 8B parameters, partially addressed by the 3B replication.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper makes no new theoretical claims. The convergence of the Kontoyiannis estimator to the true entropy rate is cited from Kontoyiannis et al. [1998]; no new proofs are offered.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 specifies all QLoRA hyperparameters (rank 16, $\alpha = 16$, learning rate 2×10^{-4}, cosine schedule, 3 epochs, effective batch size 8, 4-bit NF4 quantisation), the document counts (80 per generation, 40 selected per condition), domain and topic structure (4 domains, 10 topics, 2 documents per topic), generation-0 sharing, and the filtering criteria. The $\hat{H}_K$ formula and the 1,500-word token truncation are given in Equations (1) and Section 3.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and a per-document metadata summary (entropy scores, domain, generation, condition – no document text) sufficient to reproduce Table 1 and Figure 1 exactly are released at `https://github.com/lewismath/no-model-required`, along with a clean reference implementation of $\hat{H}_K$ (Equation (1)). The full raw document corpus (1.6 GB, all conditions and generations) is separately available at `https://huggingface.co/datasets/lewismath/phase5-raw`.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Full training configuration is given in Section 5: QLoRA rank, α, learning rate, schedule, epochs, batch size, quantisation format. Data splits (generation-0 shared; 40 of 80 documents selected per condition per generation; domain-stratified 10 per domain) and test set (80 held-out generation-6 documents per condition) are described. Model identifiers and API versions are cited or named.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All figures include ± 1 SEM bands ($n = 80$ per generation); Figure 1b uses bootstrap 95% confidence intervals; Table 1 reports Welch t-test p-values for all pairwise condition comparisons; collapse detection results include Spearman ρ with p-values and sample sizes. Error bar definitions are stated in figure captions.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper identifies the compute infrastructure (single NVIDIA A100 GPU on the Phoenix HPC cluster at the University of Adelaide, acknowledged in the Acknowledgments) and notes that QLoRA makes multi-generation fine-tuning tractable on a single A100. One run of 6 fine-tuning generations takes approximately 24 hours on this infrastructure.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research studies training-data filtering methods to reduce model collapse in synthetic fine-tuning pipelines. No human subjects are involved, no sensitive personal data are collected or processed, and no harmful applications are enabled by this work.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: A Broader Impacts paragraph appears at the end of Section 6. It identifies the primary positive impact (accessible collapse-resistant curation without model access; counteracting homogenisation of synthetic web content) and the main negative consideration (making fully-synthetic pipelines more viable may reduce incentives to collect or compensate human-authored training data).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper introduces a text-analysis algorithm ($\hat{H}_K$ filtering), not a new model or scraped dataset. No assets with high misuse potential are released.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: All models used are properly cited: Llama-3.1-8B-Instruct [Dubey et al., 2024] (Meta Llama 3 Community License), GPT-4o and GPT-2 [Radford et al., 2019] (OpenAI Terms of Service), Gemini 1.5 Pro [Gemini Team, Google, 2024] (Google Terms of Service), Qwen2.5-3B [Yang et al., 2024] (Qwen License). API usage conforms to each provider's terms of service.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released repository (`https://github.com/lewismath/no-model-required`) is documented with a README describing its structure, data policy, installation, and usage; it mirrors the paper's own section structure and includes a runnable example of the $\hat{H}_K$ estimator.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. All data are generated by language model APIs.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects are involved in any part of this research.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [Yes]

    Justification: LLMs are the objects of study and the source of all experimental data. GPT-4o and Gemini 1.5 Pro are used to generate the document corpora for proxy validation (Section 3); Llama-3.1-8B-Instruct is both the generator and the model subject to iterative fine-tuning in Sections 4 and 5; Qwen2.5-3B is used as an independent scorer for cross-model validation. All models are identified by name and cited.