Per-Document AI-Text Detection Is Not a Reliable Model-Collapse Monitor

Shaheer Saud, *Independent*, shaheersaud2004@gmail.com

Abstract

Model collapse—degradation under recursive training on generated data [Shumailov2023]—is often conflated with per-document AI-text detection. We argue these are distinct objectives: detectors that correctly flag synthetic documents can fail to register distributional degeneration. We evaluate a six-statistic heuristic (SyntheticTextProbe) and a leave-one-model-out logistic classifier over those statistics plus one perplexity feature on HC3 and RAID (eight generators). The classifier attains mean held-out AUC 0.915 ± 0.006 (10 seeds; 95% CI [0.911, 0.920]), above perplexity alone (0.903 ± 0.007; paired permutation $p = 0.002$) and a fixed zero-shot heuristic (0.732). Replacing an in-distribution bigram perplexity feature with frozen GPT-2 (124M) perplexity does not improve mean AUC under this protocol (0.895), a comparison we flag as information-mismatched. In recursive-collapse experiments (trigram LM, distilGPT-2, Qwen2.5-0.5B), vocabulary size, n-gram diversity, and reference perplexity change monotonically with collapse, while per-document detector scores remain flat or sub-threshold (e.g., ~91% of collapsed distilGPT-2 outputs unflagged). Stronger likelihood detectors (GLTR-style) improve document AUC but share the same unit-of-analysis limitation. Collapse monitoring should track corpus-level diversity and reference perplexity rather than aggregated document verdicts.

1. Introduction

Recursive training on model-generated text can induce *model collapse*: progressive loss of distributional tails and diversity [Shumailov2023] (popularized as "Habsburg AI"). In parallel, a large literature builds *per-document* detectors that score whether a given passage is machine-written. These problems are often treated as interchangeable—e.g., by aggregating detector verdicts as a contamination or collapse alarm—but they operate at different units of analysis. A detector can correctly label synthetic documents while remaining insensitive to corpus-level degeneration. This paper separates the two tasks empirically. Using six hand-crafted text statistics plus a single perplexity feature under leave-one-model-out evaluation on RAID [Dugan2024], we show that lightweight detection transfers to unseen generators (mean AUC 0.915 ± 0.006 over 10 seeds). Under controlled recursive retraining of a trigram LM, distilGPT-2, and Qwen2.5-0.5B, however, distribution-level metrics (vocabulary size, n-gram diversity, reference perplexity) track collapse monotonically, whereas the same per-document detector stays flat or below threshold. The contribution is methodological: contamination detection at document granularity is not a reliable monitor of model collapse under the settings we test.

Contributions

1. **Document-level detection is not collapse monitoring.** Across a trigram LM, distilGPT-2, and Qwen2.5-0.5B—and corpus sizes from 150 to 15,000 documents—distribution-level metrics move monotonically with collapse while the per-document detector stays flat or sub-threshold (~91% of collapsed distilGPT-2 documents unflagged).
2. **Cheap features generalize under LOMO.** Six statistics plus perplexity reach mean held-out AUC 0.915 ± 0.006 on RAID (10 seeds; paired lift over perplexity alone +0.013, $p = 0.002$).
3. **Boundary results.** Zero-shot transfer is uneven (strong on RLHF chat models, near chance on GPT-2 base); frozen GPT-2 perplexity does not beat an in-distribution bigram under mismatched information conditions; a stronger GLTR-style detector still remains per-document.

Section 2 situates the work. Sections 3–4 describe method and setup. Section 5 reports detection, then collapse. Sections 6–8 discuss implications, limitations, and conclusions.

---

2. Related Work

Model collapse. Shumailov et al. [Shumailov2023] showed that recursive training on generated data causes models to forget rare events and contract effective support. We instrument that process with cheap distribution-level statistics; we do not claim to rediscover it.

Statistical AI-text detection. GLTR [Gehrmann2019] and DetectGPT [Mitchell2023] detect machine text from likelihood or curvature under a scoring model. Our SyntheticTextProbe uses six surface statistics without white-box token probabilities, trading strength for transparency. We compare against GLTR under GPT-2 in Section 5.6.

Reliability and datasets. Sadasivan et al. [Sadasivan2023] argue that reliable detection becomes impossible as generators approach the human distribution; our uneven zero-shot transfer is consistent with that caution. We use HC3 [Guo2023] for single-generator calibration and RAID [Dugan2024] for multi-generator leave-one-model-out evaluation.

3. Method

3.1 SyntheticTextProbe

The probe maps a document to a synthetic-likelihood score in $[0, 100]$ as a fixed-weight linear combination of six normalized signals (Table 1). Human-high signals (lexical diversity, n-gram diversity, burstiness) are inverted so that every contribution increases with synthetic likelihood. Predict synthetic if score ≥ 50.

$$\text{score} = 100 \cdot \sum_{i=1}^{6} w_i \, s_i, \qquad \sum_{i=1}^{6} w_i = 1.0.$$

Weights and the threshold of 50 are hand-fixed, not tuned per dataset. Without a reference corpus, `duplicate_similarity` is defined as 0 and is inert.

3.2 Baselines

- **Bigram perplexity:** word-level bigram LM with add-k smoothing ($k = 0.5$), re-fit on an in-distribution reference split when used as a LOMO feature.
- **Mean word length:** trivial sanity floor.
- **GPT-2 (124M) perplexity:** frozen, zero-shot [Radford2019]; never fit to RAID/HC3.

3.3 Leave-one-model-out classifier

A logistic regression over **seven features** (six probe signals + one perplexity feature), with internal z-score standardization and L_2 regularization (10^{-3}). Protocol: train on seven RAID generators plus human text; test on the held-out eighth. Human partitions are fixed across folds (2000 train / 400 test, balanced, disjoint). Seed 1234 for the reference run; 10 seeds for the headline mean.

4. Experimental Setup

Datasets. HC3 [Guo2023]: 2500 human + 2500 ChatGPT answers to the same questions (~250-word truncation); 70/30 split (test $n = 1500$). RAID [Dugan2024]: attack=`none`; eight generators (gpt2, gpt3, gpt4, chatgpt, llama-chat, mistral-chat, cohere, mpt-chat) $\times$ 400 documents + 3000 human, same truncation.

Metrics. Primary: AUC (rank-based / Mann–Whitney, positive = synthetic). Also accuracy, precision, recall, F1 at the default threshold. Collapse runs additionally track vocabulary size, distinct-trigram diversity, type–token ratio, and reference-model perplexity.

Scope of statistical rigor. Multi-seed confidence intervals and a paired permutation test are reported for the headline bigram-feature LOMO result only. GPT-2-feature LOMO, cross-model tables, GLTR, and collapse trajectories are single-seed point estimates.

Compute. Core pipeline is Python standard library. An isolated PyTorch/Transformers environment is used only for GPT-2 perplexity and neural collapse (Apple MPS). Code, seeds, and result JSONs are in the repository.

5. Results

5.1 HC3 calibration

On held-out HC3, the probe reaches AUC 0.912 (Table 2), ahead of re-fit bigram perplexity (0.832) and a mean-word-length floor (0.646). This is a favorable single-generator setting and serves as calibration, not the main generalization claim. Ablation (Appendix A): burstiness and lexical diversity dominate; `duplicate_similarity` is inert without a reference corpus. ROC curves appear in Figure 1.

5.2 Uneven zero-shot transfer (RAID)

A fixed zero-shot probe transfers unevenly across RAID generators (Table 3; Figure 2): strongest on RLHF/chat models (chatgpt 0.829, llama-chat 0.794), near chance on gpt2 (0.566). Pooled AUC is 0.723 (mean across generators 0.710). The bigram column is higher but re-fit in-distribution and is not a fair zero-shot comparison.

5.3 Leave-one-model-out generalization

Under LOMO, the trained classifier reaches mean AUC **0.924** at the reference seed (Table 4; Figure 3) versus 0.911 for perplexity alone and 0.732 for the zero-shot probe. It holds on the hardest held-outs (gpt2 0.834, cohere 0.837). Across **10 seeds**, classifier mean is **0.915 ± 0.006** (95% CI [0.911, 0.920]) vs. perplexity alone

Table 1: SyntheticTextProbe signals: weight, what each measures, and orientation.

Signal	Weight	Measures	Orientation
`repetition`	0.18	Repeated-trigram fraction	synthetic-high
`filler_density`	0.20	Canned filler phrases / ~50 tokens	synthetic-high
`lexical_diversity`	0.16	Type–token ratio	human-high (inverted)
`ngram_diversity`	0.16	Distinct-trigram ratio	human-high (inverted)
`burstiness`	0.15	CV of sentence lengths	human-high (inverted)
`duplicate_similarity`	0.15	Max Jaccard vs. reference model outputs	synthetic-high

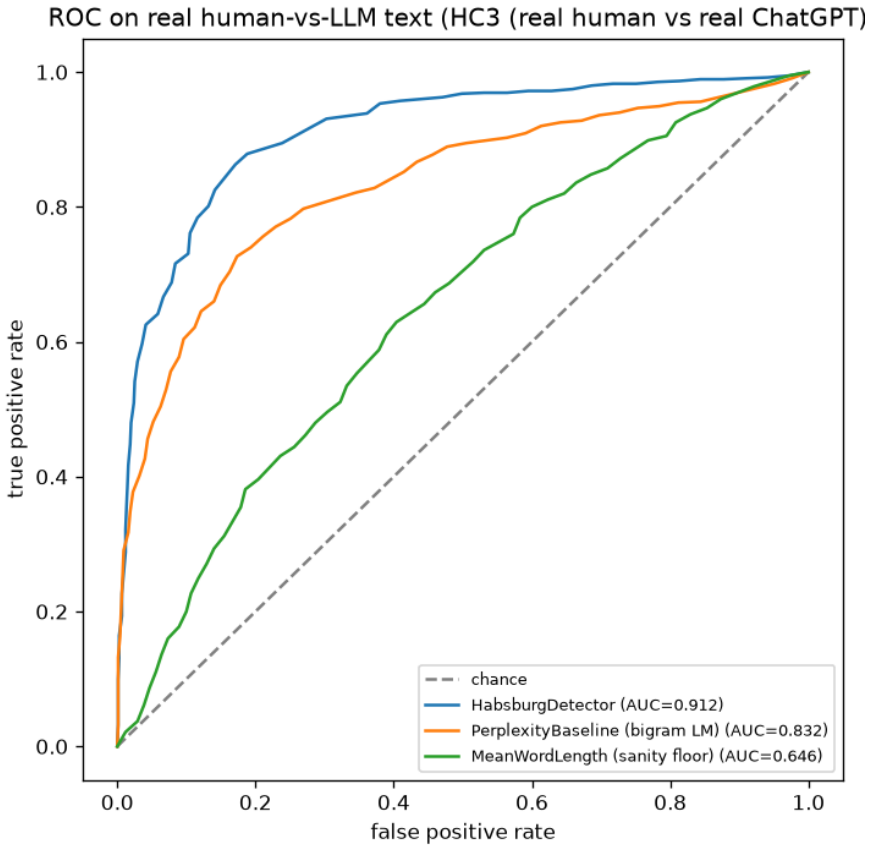


Figure 1: ROC on HC3 (real human vs. ChatGPT): SyntheticTextProbe against the perplexity and word-length baselines.

0.903 ± 0.007; mean gain **+0.013** (computed per-seed, before rounding), positive on **10/10** seeds, paired permutation $p = 0.002$. The lift is modest and concentrated where perplexity is weaker (mpt-chat +0.065, gpt4 +0.039); on gpt2, gpt3, and cohere, perplexity alone is marginally ahead.

Feature importance (Table 5): perplexity dominates; lexical diversity and repetition are secondary; `duplicate_similarity` is 0 without a reference corpus. When the perplexity feature is weakened (frozen GPT-2), the classifier leans more on the surface statistics.

5.4 GPT-2 perplexity as a feature (negative)

Replacing the in-distribution bigram with frozen GPT-2 (124M) perplexity yields mean classifier AUC **0.895** (vs. 0.924) and perplexity-alone **0.864** (vs. 0.911) at the reference seed. This is **not** an apples-to-apples capacity comparison: the bigram is re-fit per fold; GPT-2 is zero-shot. Under these mismatched information conditions, a larger LM is not automatically a better detection feature. Matched comparison is left to future work.

5.5 Model collapse: detection vs. monitoring

We recursively retrain generators on their own output and track distribution-level metrics alongside the per-document detector (seed: 150 RAID human documents unless noted). We did not discover collapse [Shumailov2023]; we ask whether a document detector tracks it.

Table 2: HC3 single-model held-out detection (positive class: ChatGPT).

Method	AUC	Acc	Prec	Rec	F1
SyntheticTextProbe	0.912	0.845	0.824	0.879	0.850
PerplexityBaseline (bigram)	0.832	0.765	0.748	0.799	0.772
MeanWordLength (floor)	0.646	0.555	0.531	0.940	0.679

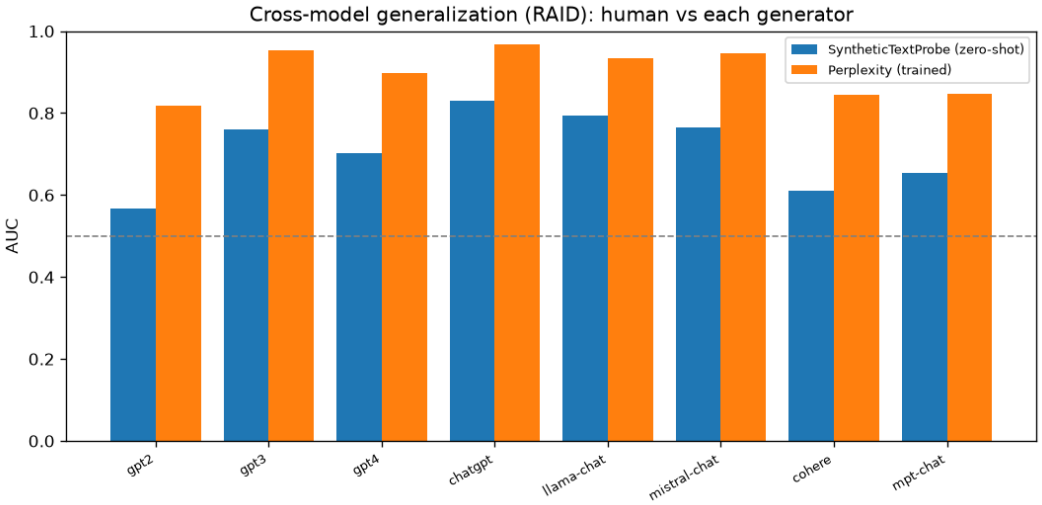


Figure 2: Cross-model detection on RAID: one fixed zero-shot detector evaluated against each of eight generators.

Summary (Table 6). Distribution metrics move monotonically; the detector does not cross its threshold of 50.

For the trigram LM, vocabulary and diversity collapse while the detector score *falls*. For distilGPT-2 [Sanh2019], reference perplexity collapses 69.2→7.4 and the detector rises 16.7→27.1 but stays below 50, leaving ~91% of collapsed documents unflagged (Figure 4). Qwen2.5-0.5B (LoRA, three generations) shows the same pattern: ref. ppl 37→2.9, detector flat.

Scale. Varying corpus size over {150, 1,500, 15,000} and seed corpus over RAID/HC3 (Figure 5): larger corpora collapse more slowly (vocab loss after five generations from −96% at 150 docs to −32% RAID / −17% HC3 at 15,000), but trigram diversity still falls (e.g. RAID 0.86→0.27 at 15k) while the detector stays flat-to-falling (~14–16 → ~9) and estimated contamination stays 0.1% (0%) in all six runs. The document-vs-distribution distinction is invariant to scale and seed corpus under these settings (Figure 6).

5.6 Stronger detectors, same limitation

Zero-shot GLTR statistics under GPT-2 [Gehrmann2019] outperform the probe (Table 7): 0.995 vs. 0.915 on HC3; 0.862 vs. 0.724 on pooled RAID. Every entry in Table 7 is a *per-document* method. The same likelihood signal becomes a collapse monitor only when computed as *corpus* reference perplexity (Section 5.5). Detector strength does not remove the unit-of-analysis gap.

6. Discussion

Lightweight detection transfers under LOMO, but the lift over perplexity is small and perplexity is the workhorse feature. Zero-shot heuristics fail where the canned, low-burstiness profile is absent (gpt2 base). The sharper implication is for monitoring practice: aggregating per-document detector scores is a poor proxy for recursive distributional degeneration. Corpus monitors should track vocabulary size, n-gram diversity, and reference perplexity over time—not mean detector score (Figure 6).

Three negative findings are part of the contribution: (i) fixed heuristics transfer unevenly; (ii) a larger frozen LM is not automatically a better feature than a domain-fitted bigram under mismatched fit conditions; (iii) a competent per-document detector is not, in our setups, a reliable collapse monitor.

7. Limitations

- **Model scale.** Collapse experiments use a trigram LM, distilGPT-2, and Qwen2.5-0.5B—not frontier models or full RLHF loops.

Table 3: Cross-model zero-shot detection on RAID, per generator. Perplexity is re-fit in-distribution and shown for reference only.

Generator	Probe (zero-shot)	Perplexity (in-dist. ref.)
gpt2	0.566	0.819
gpt3	0.759	0.953
gpt4	0.703	0.897
chatgpt	0.829	0.968
llama-chat	0.794	0.934
mistral-chat	0.764	0.945
cohere	0.610	0.845
mpt-chat	0.654	0.847
Pooled	0.723	0.908

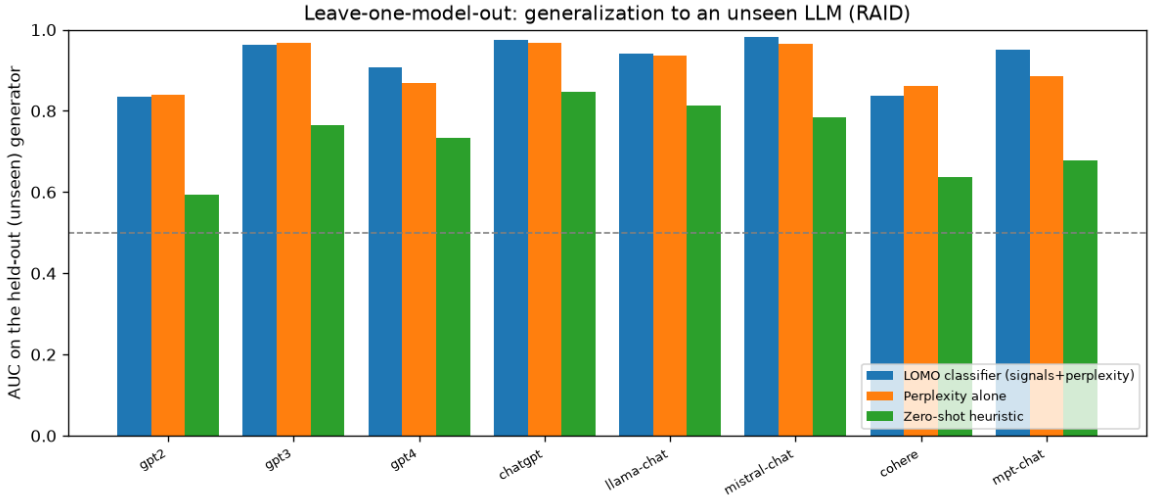


Figure 3: Leave-one-model-out: AUC on the held-out (unseen) generator for the trained classifier, perplexity alone, and the zero-shot heuristic.

- **Language and data.** Features and corpora are English; RAID/HC3 are dated; texts are ~250-word excerpts.
- **Proxies.** Vocabulary, diversity, and reference perplexity proxy collapse; factual drift, bias, and long-range incoherence are unmeasured.
- **Decoding.** Temperature / top-p / repetition penalty are fixed; other sampling can change magnitudes.
- **Statistics.** Multi-seed significance applies only to headline LOMO; other results are point estimates.
- **Perplexity comparison.** Bigram is in-distribution; GPT-2 is frozen—information conditions differ.
- **Untuned probe.** Weights and threshold 50 are hand-fixed; `duplicate_similarity` is inert here.

We claim a qualitative distinction under the settings above, not universal quantitative collapse rates. A stronger document detector does not automatically become a distributional monitor (Section 5.6).

8. Conclusion

Cheap statistics plus perplexity generalize to unseen RAID generators (LOMO AUC 0.915 ± 0.006). A larger frozen LM is not automatically a better feature under mismatched fit. Most importantly, under recursive retraining, distribution-level metrics track collapse while per-document detection stays flat or sub-threshold. Detection asks whether *this* text is synthetic; monitoring asks whether the *distribution* is narrowing—and they require different signals.

Future work. (1) Matched perplexity comparisons and multi-seed analysis beyond headline LOMO. (2) Dedicated corpus-level monitors (e.g., embedding-space diversity, semantic entropy) on larger mixed human–synthetic corpora.

References

[Dugan2024] L. Dugan et al. “RAID: A Shared Benchmark for Robust Evaluation of Machine-

Table 4: Leave-one-model-out AUC on the unseen generator (six signals + bigram perplexity). Reference seed 1234; 10-seed mean classifier AUC = 0.915 ± 0.006. The zero-shot column is recomputed on the LOMO held-out human partition and therefore differs from Table 3.

Held-out	Classifier	Perplexity-alone	Zero-shot
gpt2	0.834	0.840	0.594
gpt3	0.962	0.967	0.765
gpt4	0.907	0.868	0.734
chatgpt	0.976	0.968	0.848
llama-chat	0.942	0.936	0.814
mistral-chat	0.981	0.965	0.784
cohere	0.837	0.862	0.636
mpt-chat	0.950	0.885	0.679
Mean	**0.924**	0.911	0.732

Table 5: LOMO feature importance: mean absolute standardized coefficient across eight folds (bigram vs. GPT-2 perplexity feature).

Feature	Bigram ppl	GPT-2 ppl
perplexity	2.73	2.49
lexical_diversity	1.24	1.46
repetition	0.58	0.90
burstiness	0.29	0.61
filler_density	0.14	0.38
ngram_diversity	0.13	0.39
duplicate_similarity	0.00	0.00

Generated Text Detectors." *Proceedings of ACL*, 2024. arXiv:2405.07940.

[Gehrmann2019] S. Gehrmann, H. Strobelt, A. M. Rush. "GLTR: Statistical Detection and Visualization of Generated Text." *Proceedings of ACL (System Demonstrations)*, 2019.

[Guo2023] B. Guo et al. "How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection." arXiv:2301.07597, 2023. (HC3 dataset.)

[Mitchell2023] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn. "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature." *Proceedings of ICML*, 2023. arXiv:2301.11305.

[Radford2019] A. Radford et al. "Language Models are Unsupervised Multitask Learners." OpenAI Technical Report, 2019. (GPT-2.)

[Sadasivan2023] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, S. Feizi. "Can AI-Generated Text be Reliably Detected?" arXiv:2303.11156, 2023.

[Sanh2019] V. Sanh, L. Debut, J. Chaumond, T. Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv:1910.01108, 2019. (distilGPT-2 lineage.)

[Shumailov2023] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, R. Anderson. "The Curse of Recursion: Training on Generated Data Makes Models Forget." arXiv:2305.17493, 2023. (Journal version: "AI models collapse when trained on recursively generated data," *Nature* 631, 2024.)

Appendix A: Ablation and Controlled Harness

HC3 per-signal ablation (full AUC 0.912). Leave-one-out dAUC: burstiness 0.037, lexical_diversity 0.036, repetition 0.012, ngram_diversity 0.004, filler_density 0.003, duplicate_similarity 0.000 (inert without reference corpus). Solo AUCs: lexical 0.848, ngram 0.850, repetition 0.848, burstiness 0.735, filler 0.613, duplicate 0.500.

Controlled sanity harness. The repository includes a manufactured corpus separable on exactly the six probe axes (near-ceiling accuracy). Those numbers are tautological and are not cited as evidence; all claims rest on HC3/RAID and the collapse experiments above.

Table 6: Collapse endpoints: distribution metrics vs. per-document detector score.

Setting	Δ vocab	Δ trigram div.	Δ ref. ppl	Detector	Est. flagged
Trigram (g0→g6)	6707→1287 (−81%)	0.943→0.270	—	16.7→11.9	0%→~1%
distilGPT-2 (g0→g5)	6707→1012 (−85%)	0.943→0.639	69.2→7.4	16.7→27.1	→~9%
Qwen2.5-0.5B (g0→g3)	−87%	—	37→2.9	17.6→18.4	2%

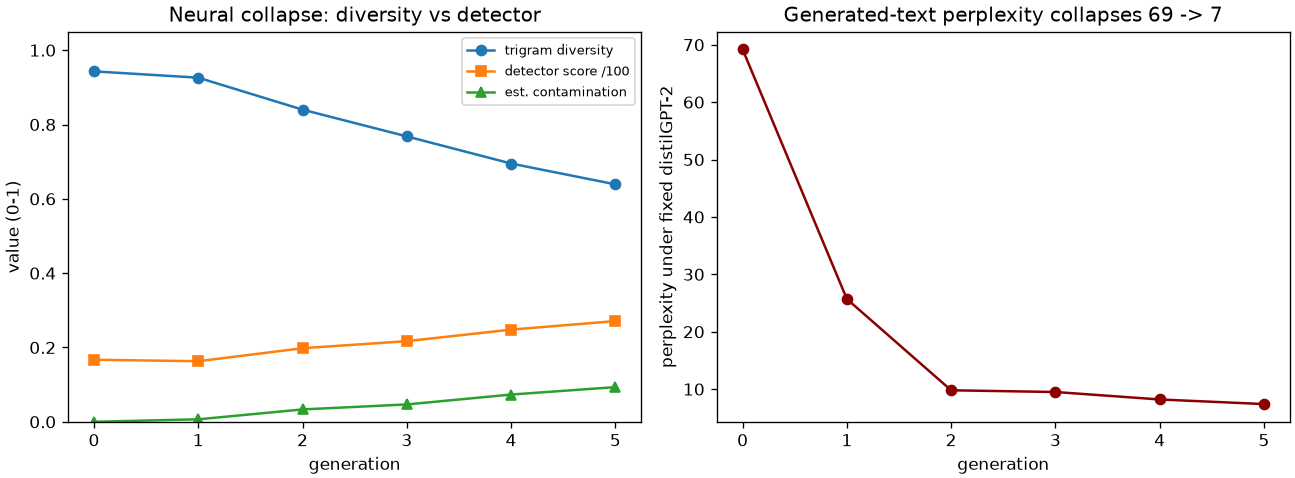


Figure 4: Neural collapse (distilGPT-2 retrained on its own output). Left: diversity and detector score per generation. Right: perplexity of generated text under a fixed reference model, 69 to 7.

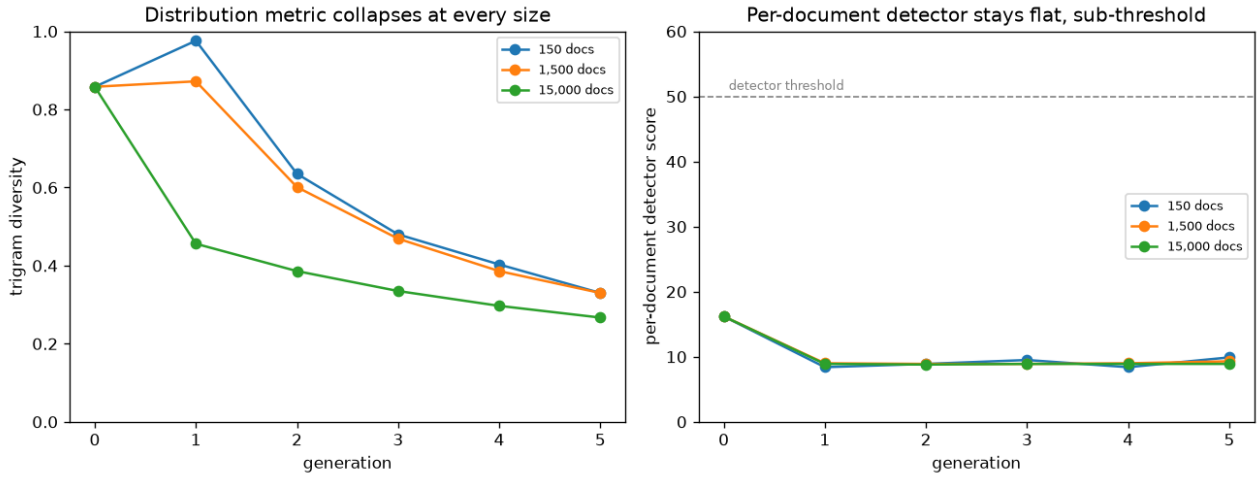
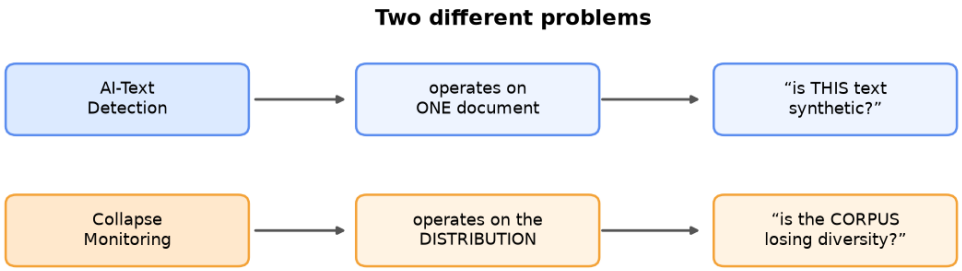


Figure 5: Collapse at scale (RAID seed; HC3 is similar). Left: trigram diversity falls at every per-generation corpus size (150 / 1,500 / 15,000 documents); larger corpora collapse more slowly but always collapse. Right: the per-document detector score stays flat and far below its threshold (dashed) at every size. The detection-vs-monitoring distinction is invariant to two orders of magnitude of scale.

Table 7: Zero-shot AUC: GLTR under GPT-2 vs. SyntheticTextProbe, computed on the full HC3/RAID pools (probe values therefore differ slightly from Tables 2–3).

Detector	HC3	RAID (pooled)
SyntheticTextProbe	0.915	0.724
GLTR — GPT-2 log-probability	0.995	0.862
GLTR — GPT-2 top-10 rank	0.995	0.848



Which signal answers which question

Signal / metric	Flag one AI document	Detect corpus collapse
Perplexity (under a fixed LM)	Good	Good
Per-document detector score	Good	Poor
Vocabulary size	Poor	Excellent
Distinct n-gram diversity	Poor	Excellent

Figure 6: The paper's central claim in one figure. AI-text detection operates on a single document; collapse monitoring operates on the distribution. Per-document statistics flag individual AI documents but miss collapse, while distribution-level statistics (vocabulary, n-gram diversity) detect collapse but say little about a single document.