

The Internet Is Becoming AI-Generated: What Happens When AI Starts Training on AI?

A review of the published evidence on synthetic data, recursive training, and model collapse — and what it does and does not establish about the future of AI training data.

Serideo Team · Nikhil V.

Serideo AI — Data Infrastructure Research

Received: August 2026 | Status: Public research summary | Version 1.1

ABSTRACT

Generative AI systems now produce a measurable share of newly published text on the open web, raising the question of what happens to future AI systems trained on data that includes prior AI output. We review two primary sources that test this directly: Shumailov et al. (*Nature*, 2024), who show that models trained recursively on their own synthetic output, with real data discarded at each step, exhibit measurable degradation they term “model collapse”; and Kazdan et al. (ICML / PMLR 267, 2025), who show that when real data is instead preserved and synthetic data simply accumulates alongside it, the same task settings do not collapse. We find the evidence supports a specific, narrower claim than “synthetic data is harmful”: outcomes depend on training-workflow design — whether real data is displaced or retained — and on the ratio of real to synthetic data present, not on the presence of synthetic data per se. We also review independent web-measurement studies on the growing share of AI-authored content, and discuss data provenance as a resulting practical concern. Every figure and quantitative claim in this report is tied to a cited source; where a source reports a trend only graphically, we say so rather than estimate a number.

CONTENTS

00 Executive Summary	2	06 Synthetic Data Is Not the Enemy	7
01 The Internet Is Changing	2	07 The Most Important Finding	10
02 A New Feedback Loop	3	08 The Data Provenance Problem	12
03 What Is Model Collapse?	4	09 Why Trusted Data Could Become More Valuable	12
04 What the Experiments Actually Showed	5	10 What Happens Next?	13
05 The Rare-Information Problem	6	11 Conclusion & References	14

KEYWORDS: model collapse · synthetic data · recursive training · data provenance · generative AI · training-data quality

00 Executive Summary

Generative AI systems now produce a meaningful share of newly published text on the open web. Independent measurement studies find that the share of sampled web articles classified as primarily AI-generated rose sharply after ChatGPT's November 2022 launch and has plateaued near 50% since early 2025 [1]. This raises a legitimate infrastructure question: what happens to future AI systems if they are trained, even partly, on data that earlier AI systems produced?

Two peer-reviewed research efforts have investigated this directly. Shumailov et al., published in *Nature* in 2024, showed that language models, Gaussian models, and other generative models degrade — a phenomenon they term **model collapse** — when each generation of a model is trained exclusively on synthetic output from the generation before it, with the original real data discarded [2]. A 2025 follow-up by Kazdan et al. (ICML / PMLR) tested the same task settings under a different, arguably more realistic assumption: that real data is *never* deleted and synthetic data simply accumulates alongside it. Under that assumption, collapse did not occur in any of the five settings they tested [3].

The evidence, read together, does not support the claim that synthetic data is inherently harmful, nor that AI model quality is on an inevitable path to collapse. It supports a narrower, more specific claim: **the training workflow — specifically, whether real data is preserved or discarded — determines the outcome much more than the mere presence of synthetic data does.** The evidence also shows a genuine interaction between real-data volume and synthetic-data value: synthetic data can measurably help when real data is scarce, and measurably hurt when real data is already abundant [3].

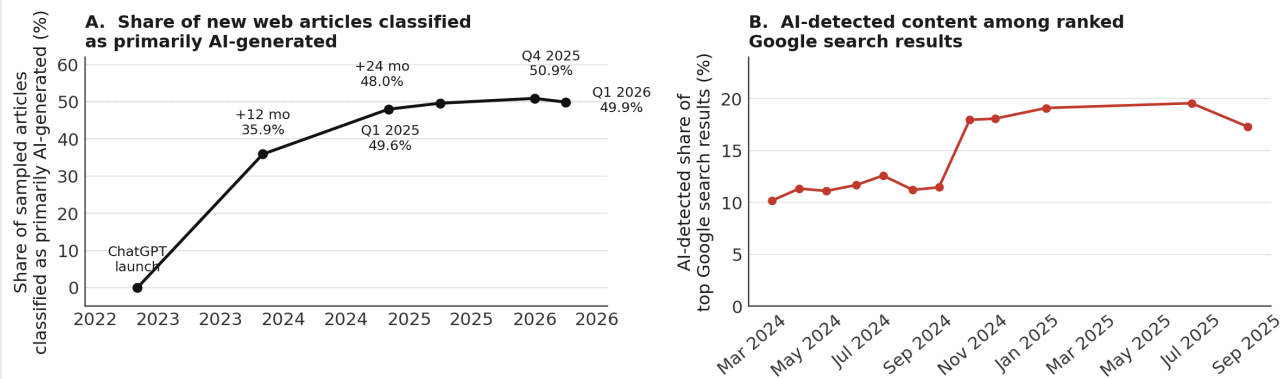
This report lays out what the primary sources actually show, what they do not show, and why data provenance — knowing where training data came from — is emerging as a practical infrastructure concern independent of whether "collapse" itself ever materializes at internet scale.

01 The Internet Is Changing

Before ChatGPT's public launch in November 2022, the overwhelming majority of new text published on the web could be safely assumed to be human-written. That assumption no longer holds uniformly. The most methodologically transparent longitudinal measurement available is a 2026 study by Graphite's "Five Percent" research team, which sampled 55,400 English-language web articles from Common Crawl — one of the largest open web archives, used widely to train language models — published between January 2020 and March 2026, and classified each using three independent AI-detection tools (Pangram, Copyleaks, and GPTZero) [1].

That study found the share of sampled articles classified as primarily AI-generated reached 35.9% within twelve months of ChatGPT's release, 48.0% within twenty-four months, and has since plateaued in a narrow band around 49–51% across five consecutive quarters through Q1 2026 [1]. A separately conducted, methodologically distinct study by Originality.ai — which audits AI-detected content specifically among top-ranking Google search results, a narrower and different population — found a similar directional trend, with AI-detected content among ranked results climbing from roughly 10% in March 2024 to roughly 18–19% by early 2025 [4].

CHART 1 — AI-Authorship Signals on the Web Over Time



Source (A): Graphite/Five Percent, "AI Now Writes as Many Online Articles as Humans" (May 2026), a random sample of 55,400 English-language articles (article/listicle schema, ≥ 100 words) drawn from Common Crawl, published Jan 2020–Mar 2026, classified with three AI detectors (Pangram, Copyleaks, GPTZero; reported false-positive rates 1.4–1.8%). Points shown are the exact figures stated in the study text; the path between points is not published and is interpolated for readability only.

Source (B): Originality.ai, "Amount of AI Content in Google Search Results" (ongoing study), an independent AI-detector-based audit of AI content appearing among top Google search rankings — a different population (ranked search results, not all published articles) and a different measurement method than (A). The two panels are not directly comparable.

Caveat: AI-detector accuracy is contested industry-wide; both studies report their own validation but detector-based estimates carry non-trivial error, and neither measures the full internet.

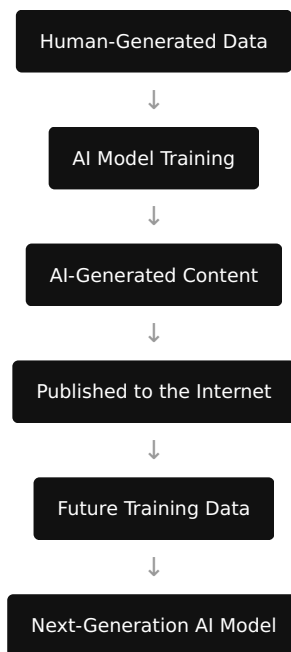
Figure 1. Two independent, methodologically distinct measurements of AI-authorship signals on the web. Panel A tracks the share of a random Common Crawl article sample classified as primarily AI-generated; Panel B tracks a different population — AI-detected content specifically among top Google search rankings. See sources for full methodology and detector validation figures. [1, 4]

Two things are worth stating precisely, because they are easy to overstate. First, neither study claims to measure "the internet" as a whole — both measure a specific, defined sample (Common Crawl articles meeting inclusion criteria; top Google search results) using AI-detection classifiers that are not perfect. The Graphite study reports its own detector false-positive rates of 1.4–1.8% and false-negative rates below 2% on recent frontier models, and is transparent that detector accuracy may degrade as models improve [1]. Second, "primarily AI-generated" does not distinguish between fully AI-authored content and AI-assisted content that a human edited or curated — a distinction the Graphite study explicitly flags as a limitation it did not evaluate [1].

What the evidence does establish is directional and significant: AI-generated content is not a marginal phenomenon in the newly published web corpus that future models will plausibly draw on. Whether that changes the composition of already-existing training archives — most of which were assembled before this shift began — is a separate, less-measured question this report returns to in Section 8.

02 A New Feedback Loop

The concern this raises is structural, not moral: AI models are trained on data scraped from the web; a growing share of that web is now itself AI output; if future models are trained on web snapshots that include this AI output, a feedback loop is created in which each generation of models is shaped, in part, by the outputs of the generation before it.



This loop, by itself, is only a description of a mechanism — it says nothing yet about whether the mechanism is harmful. Whether it degrades future models depends entirely on *how* that future training data is assembled: whether AI-generated text sits alongside a large, stable pool of human-generated data, or whether it substitutes for it. This is precisely the distinction the research in Sections 3–7 is designed to test.

It is also worth being precise about what is and is not established here. The loop above describes a plausible pathway, documented in the literature as a subject of active research [2, 3]. It is not itself a measured outcome — no study in the evidence base directly tracks a real-world model's performance degrading over multiple internet-scale training generations, because no organization has run that experiment at that scale. The evidence instead comes from controlled experiments — described next — that isolate the mechanism under laboratory conditions.

03 What Is Model Collapse?

Shumailov et al. define model collapse as "a degenerative process affecting generations of learned generative models, in which the data they generate end up polluting the training set of the next generation. Being trained on polluted data, they then mis-perceive reality." [2]

In plain terms: imagine photocopying a photocopy, repeatedly. Each copy is made not from the original document but from the previous copy. Small imperfections in each copy — a slightly faded line, a blurred edge — get copied into the next generation along with everything else. After enough generations, the fine detail is gone, and what's left is a rough, blurry approximation of the original that has actually drifted in specific, non-random ways.

The paper distinguishes two stages. In **early model collapse**, a model starts losing information about the rare, unusual, or low-probability parts of its training data — the "tails" of the distribution — while the common, typical patterns are preserved and even reinforced. In **late model collapse**, the process has gone further: the model's output has drifted so far from the original data that it bears little resemblance to it at all, typically collapsing toward a narrow, repetitive range of outputs [2].

Critically, this happens even under close-to-ideal conditions. The paper's mathematical analysis (Theorem 3.1, for the Gaussian case) proves that even with a fixed sample size and unbiased statistical estimators — that is, even when nothing is going obviously "wrong" with the fitting procedure — repeatedly re-estimating a distribution from its own

generated samples causes the estimated variance to collapse toward zero and the estimated distribution to diverge, in the mathematical (Wasserstein-2) sense, from the true original distribution as the number of generations grows [2]. Collapse in this framework is not a bug in a particular training pipeline; it is a mathematical property of the resampling process itself, when nothing external replenishes the loop.

04 What the Experiments Actually Showed

Shumailov et al.'s central language-model experiment fine-tuned OPT-125m, a 125-million-parameter causal language model released by Meta, on the WikiText-2 dataset [2]. The model fine-tuned directly on real WikiText-2 data achieved a mean perplexity of 34, a large improvement over the 115 perplexity of the same model with no fine-tuning at all (perplexity is a standard language-modeling metric where lower values indicate the model assigns higher probability to — is less "surprised" by — real held-out text) [2].

From that fine-tuned starting point, the authors generated a synthetic dataset by having the model complete text continuations (via five-way beam search, in blocks of 64 tokens) for every sequence in the original training set, then fine-tuned a new "generation" of the model on that synthetic output, and repeated this for nine generations, under two conditions: one in which no original real data was retained after generation zero, and one in which a random 10% of the original real data was resampled and mixed in at every generation [2].

KEY FINDING

Zero-shot perplexity: 115 → Fine-tuned on real data: 34

Source: Shumailov et al., Nature, 2024 — stated directly in the Results text.

KEY FINDING

With no original data retained across nine generations, performance loss of 20-28 perplexity points was reported

Source: Shumailov et al., Nature, 2024. The 10%-data-preserved condition is reported only qualitatively as producing "only minor degradation of performance" — no comparable exact figure is published in the article text.

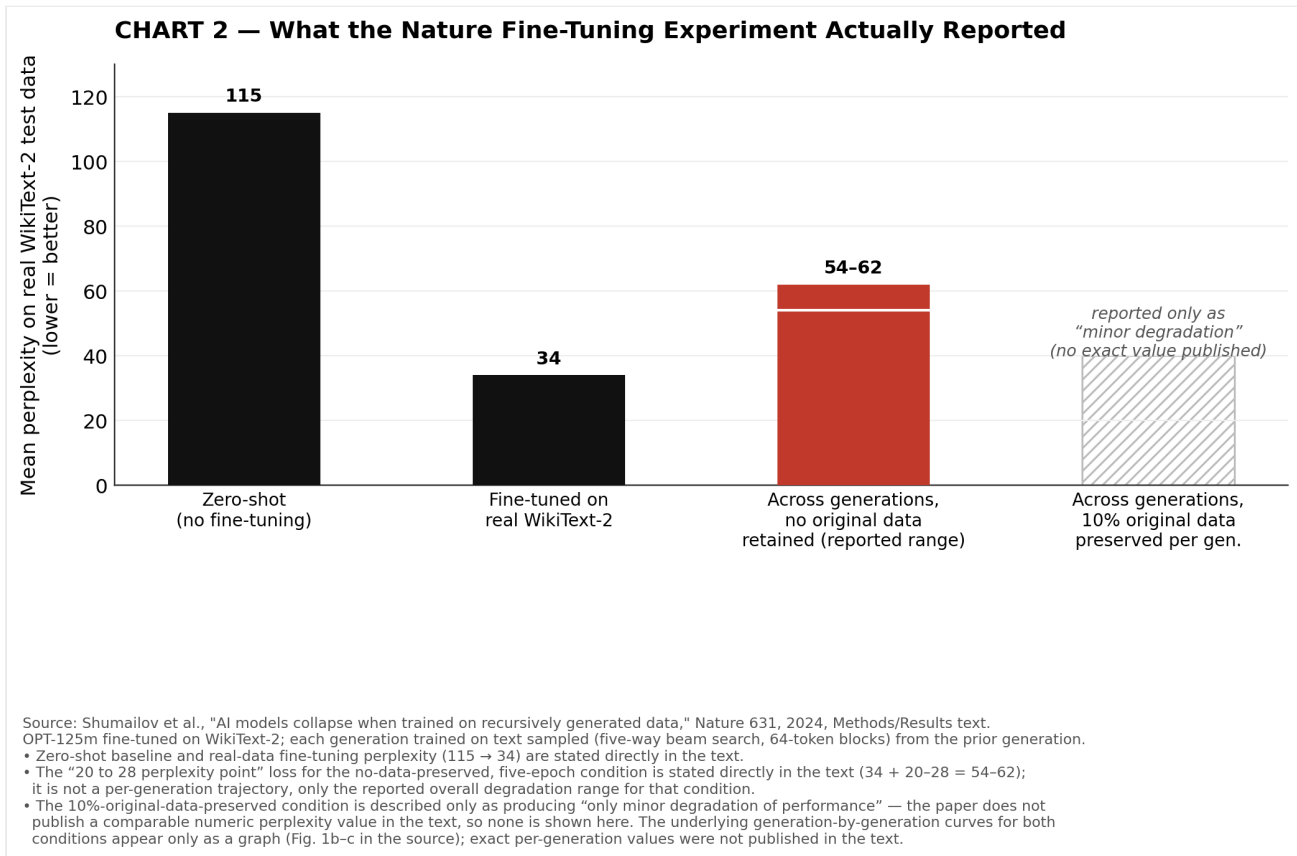


Figure 2. The exact anchor values the Nature paper states in its text. The paper's Figure 1b-c additionally plots full generation-by-generation perplexity curves and distribution histograms for both conditions — that generation-by-generation trend is reported graphically only; exact underlying per-generation values were not published in the article text, so they are not reproduced as numbers here.

Both conditions showed degraded performance relative to the real-data baseline — but the paper is explicit that the degree of degradation differed substantially: the no-preservation condition showed clear, worsening degradation across generations, while the 10%-preservation condition showed comparatively minor degradation [2]. This single result — that retaining even a modest fraction of original data materially changes the outcome — is the empirical seed that the 2025 Kazdan et al. follow-up study builds directly on (Section 6).

The paper also ran a qualitative check often cited in popular coverage: it printed example text outputs from generations 0, 1, 5, and 9 of the no-preservation condition, fine-tuned on a passage about medieval English church architecture. By generation 9, the model's continuation of that same prompt had drifted into a list of jackrabbit subspecies and their coloring — illustrating, anecdotally, the "late collapse" pattern of output narrowing into a repetitive, disconnected range unrelated to the original topic [2]. The authors are careful to note this is an illustrative example, not a quantitative result.

05 The Rare-Information Problem

The mechanism behind early model collapse, as the Nature paper's theoretical section describes it, is specifically about the loss of low-probability information. When a model is trained on a finite sample drawn from a distribution, states or patterns that occur below a certain threshold frequency have a non-trivial chance of simply not appearing in that finite sample at all — and once absent from a generation's training data, they cannot reappear in later generations, because later generations only ever sample from what earlier generations produced [2]. The paper proves this formally for a

discrete-distribution case: the process is mathematically a Markov chain whose only possible long-run resting points (absorbing states) are distributions concentrated on a single value — the model has, in the limit, forgotten everything except one outcome [2].

Why does this matter beyond the mathematics? The paper's Discussion section makes the connection explicit: "Preserving the ability of LLMs to model low-probability events is essential to the fairness of their predictions: such events are often relevant to marginalized groups. Low-probability events are also vital to understand complex systems." [2] In practical terms, the rare tail of a data distribution is disproportionately where unusual dialects, minority perspectives, edge-case scenarios, and atypical-but-important information live. A model that has quietly lost its grip on the tail of its training distribution does not necessarily look obviously broken on average — it can still perform well on common, typical inputs — while becoming systematically less reliable on the unusual ones.

This is a claim the source directly supports; it is not, however, a claim that has been independently verified at the scale of a deployed, general-purpose commercial system. The Nature paper's demonstration is on a 125-million-parameter model fine-tuned on a specific 2-million-token dataset — informative about the underlying mechanism, but not itself a measurement of rare-information loss in any current production language model.

06 But Synthetic Data Is Not the Enemy

The Nature paper's own experiments already hinted that *how* synthetic data is used matters as much as whether it's used at all — the 10%-real-data-preserved condition degraded far less than the zero-preservation condition. The 2025 ICML paper by Kazdan, Schaeffer, Dey, Gerstgrasser, Rafailov, Donoho, and Koyejo (Stanford) takes this observation and tests it directly and formally [3].

The authors define and compare three distinct **training-workflows** — precise procedural definitions of how data is assembled at each generation, not properties of the data itself:

- **Replace** — the workflow Shumailov et al. primarily studied: at each generation, all prior data is deleted and the next model trains only on data freshly generated by the previous model.
- **Accumulate** — real data is never deleted; each new generation of synthetic data is added to the pool, and every model trains on the full accumulated set (real plus all synthetic generations so far).
- **Accumulate-subsample** — a more computationally realistic middle ground: data accumulates as above, but because training compute doesn't grow to match, each generation is trained on a fixed-size sample drawn from the growing pool, rather than the whole thing [3].

The paper tests all three workflows across five separate generative-modeling task settings: multivariate Gaussian modeling, kernel density estimation, linear regression, supervised fine-tuning of language models, and language-model pretraining [3]. The result, consistent across every setting tested: **replace reliably produces collapse; accumulate does not.**

CHART 4 — Two Different Ways Synthetic Data Can Enter Training (Schematic)

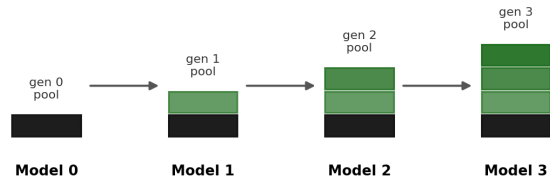
A. “Replace” workflow

Each generation **DELETES** the prior data
and trains only on its own synthetic output



B. “Accumulate” workflow

Each generation **KEEPS** all prior real + synthetic data
and trains on the growing combined pool



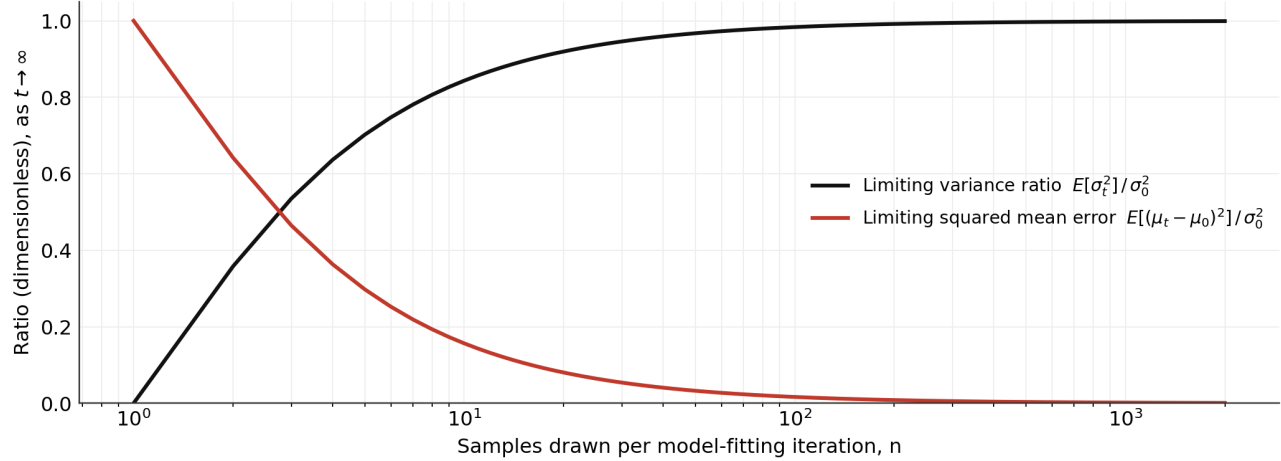
This is an illustrative schematic of the two training-workflow definitions used in Shumailov et al. (Nature, 2024) and Kazdan et al. (ICML, 2025) — it does not plot experimental data. Both papers report (via graphs, not tables of per-generation numbers) that:

- Replace: real data is deleted after generation 0 and each later model trains only on the previous model's output — both papers find this induces model collapse across every task setting they test (multivariate Gaussian modeling, kernel density estimation, linear regression, and language-model fine-tuning/pretraining).
- Accumulate: real data is never deleted; each generation trains on real data plus every synthetic generation produced so far — Kazdan et al. report this avoids collapse (test loss stays bounded) in the same task settings, though a more realistic fixed-compute variant ('accumulate-subsample', not shown here) still shows slow, bounded degradation rather than full collapse.

Figure 3. Schematic illustration of the two workflow definitions (this diagram illustrates the definitions; it does not plot experimental results — see Figure 4 and Figure 5 for the paper's actual measured outcomes).

For the multivariate Gaussian case, the authors go further and prove a closed-form asymptotic result (their Theorem 1): under accumulate, as the number of generations grows without bound, both the estimated variance and the squared error of the estimated mean converge to a *finite* limit that depends on the per-generation sample size — rather than the variance collapsing to zero and the distributional distance diverging to infinity, as Shumailov et al. proved occurs under replace [3].

**CHART 3 — Under the “Accumulate” Workflow, Error Stays Bounded
(closed-form result, univariate Gaussian case)**



Source: Kazdan, Schaeffer, Dey, Gerstgrasser, Rafailov, Donoho & Koyejo, “Collapse or Thrive? Perils and Promises of Synthetic Data in a Self-Generating World,” ICML / PMLR 267, 2025 — Theorem 1 (Appendix B). Curves are the paper’s own closed-form asymptotic ($t \rightarrow \infty$) expressions for the univariate-Gaussian ‘accumulate’ workflow, evaluated here across a range of per-iteration sample sizes n ; they are not simulation output. The key qualitative result the paper draws from this: under accumulate, both quantities converge to a finite bound (neither the variance collapses to 0 nor the mean error diverges), unlike the ‘replace’ workflow, where the paper reports (citing Shumailov et al. 2024) that variance provably collapses to 0 and Wasserstein-2 distance from the true distribution provably diverges as generations $\rightarrow \infty$.

Figure 4. The paper’s own closed-form mathematical result for the accumulate workflow (univariate Gaussian case), plotted across a range of per-generation sample sizes. As the per-iteration sample size n grows, the accumulate workflow’s limiting variance ratio approaches 1 (variance is preserved) and its limiting mean-error ratio approaches 0 (the estimated mean does not drift) — the opposite of the replace workflow’s proven behavior.

Their supervised fine-tuning experiment used Google’s Gemma 2 models (2B, 9B, and 27B parameters) fine-tuned iteratively on Nvidia’s HelpSteer2 instruction-following dataset. Under replace, each new model trained only on the roughly 12,500 synthetic examples generated by the model before it; under accumulate, each new model trained on the original real data plus every synthetic generation produced so far, so the training set grew by roughly 12,500 examples per generation [3]. The paper reports that replace produced clear collapse (rising cross-entropy loss on real held-out data across generations) for all three model sizes, while accumulate did not [3].

KEY FINDING

Across three model sizes (Gemma 2 2B / 9B / 27B) and five task settings, the accumulate workflow avoided the collapse seen under replace

Source: Kazdan et al., ICML / PMLR 267, 2025 — Sections 2 and 3, Figures 1–4.

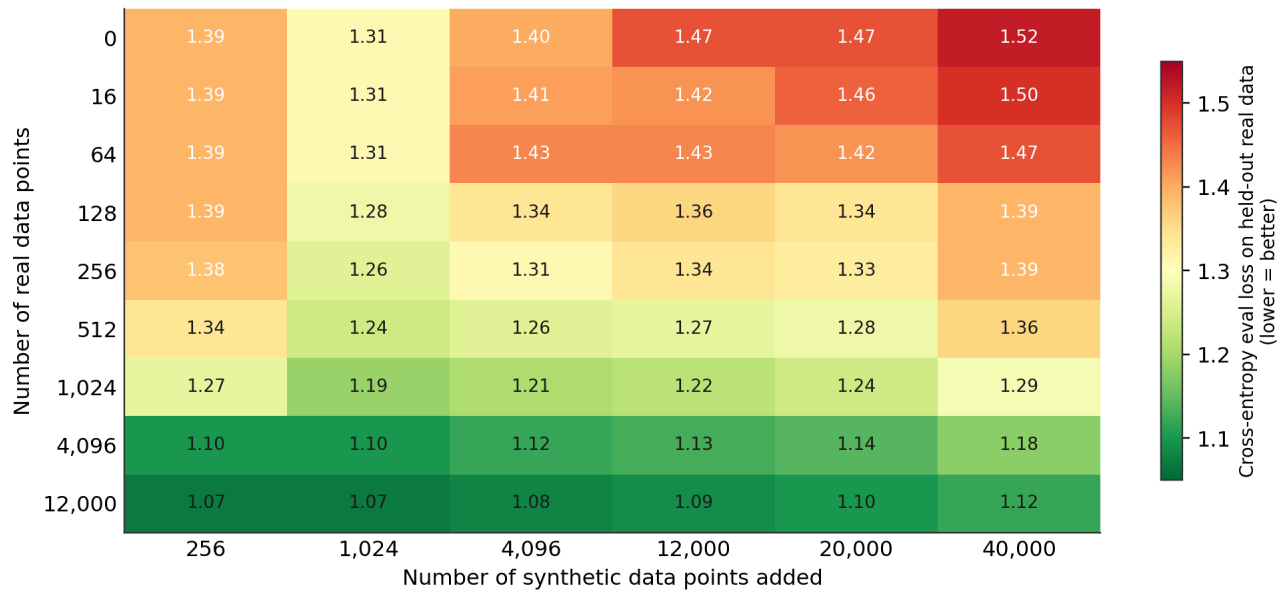
07 The Most Important Finding

The problem may not be synthetic data itself. The deeper problem may be replacing the original distribution with recursively generated data — not generating synthetic data in the first place.

This distinction matters because it points to different practical responses. If the core risk were "synthetic data is contaminating," the natural (if impractical) response would be trying to eliminate AI-generated content from training pipelines altogether. If the core risk is instead "real data is being displaced rather than supplemented," the practical response looks different: preserve, expand, and be able to identify verified real/human-originated data, and treat data-workflow design — not data-source purity — as the actual point of control.

The Kazdan et al. paper strengthens this reading with a further, more granular result: fine-tuning Gemma 2 2B on every combination of real-data volume (0 to 12,000 examples) and synthetic-data volume (0 to 40,000 examples) they tested, using 100,000 model-generated completions filtered down to 55,000 usable samples, produced a clear and statistically significant interaction, not a simple "synthetic data is bad" pattern [3].

**CHART 5 — Real vs. Synthetic Data: Measured Eval Loss
(Gemma 2 2B, fine-tuned on HelpSteer2 + generated completions)**



Source: Kazdan, Schaeffer, Dey, Gerstgrasser, Rafailov, Donoho & Koyejo, "Collapse or Thrive? Perils and Promises of Synthetic Data in a Self-Generating World," ICML / PMLR 267, 2025, Figure 5 (exact table of eval-loss values as printed in the source). Google's Gemma 2 2B was fine-tuned on Nvidia's HelpSteer2 dataset using varying mixes of real data and model-generated synthetic completions (100,000 completions generated, filtered to 55,000 usable samples ≤ 512 tokens). Reading the grid: loss falls (greener) moving down any column — more real data helps — but at low real-data counts (top rows), adding a moderate amount of synthetic data (left-of-center columns) lowers loss versus adding none; at high real-data counts (bottom rows), adding more synthetic data tends to raise loss slightly. The paper reports this as a genuine interaction, not just a real-data-volume effect ($R^2 = 0.59$ for a transformed real-data-count term, $R^2 = 0.34$ for real-data proportion; both significant at $p < 10^{-24}$ by their F-test).

Figure 5. Exact eval-loss values reported in Kazdan et al.'s Figure 5, reproduced as an original chart. Reading down any column, more real data reliably reduces loss. But reading across a row, the effect of adding synthetic data depends on how much real data is already present: at the top rows (real data scarce), a moderate amount of synthetic data reduces loss relative to none; at the bottom rows (real data abundant), added synthetic data tends to raise loss slightly.

KEY FINDING

100,000 generated completions were filtered down to 55,000 usable samples for this experiment

Source: Kazdan et al., ICML / PMLR 267, 2025, Section 4.

KEY FINDING

$R^2 = 0.59$ for a transformed real-data-cardinality term; $R^2 = 0.34$ for real-data proportion (both $p < 6.9 \times 10^{-25}$)

Source: Kazdan et al., ICML / PMLR 267, 2025, Section 4. The authors conclude both the raw amount and the proportion of real data have a statistically significant, independent effect on final test loss.

The paper's own summary of this finding is directly quotable and worth stating plainly: when real data is scarce (roughly 1,024 examples or fewer, in their experiment), adding some synthetic data reduced test loss, with the improvement peaking near 1,024 added synthetic examples. When real data was already abundant (more than roughly 1,024 examples), adding synthetic data on top of it "almost always" increased test loss, and datasets containing only real data outperformed datasets combining the same real data with roughly ten times as much synthetic data added [3].

08 The Data Provenance Problem

Put Sections 1 and 6–7 together and a practical question follows directly: if training outcomes depend on whether data is real or synthetic, and on whether real data is being preserved or displaced, then knowing which is which — for any given piece of data flowing into a training pipeline — becomes operationally important. This is the data provenance problem: for any given document, image, or dataset entering a training pipeline, can you answer:

- Where did this data originate?
- Was any part of it AI-generated, and by which system?
- Has a human reviewed, edited, or verified it?
- Has this data (or data derived from it) already been used to train a prior model?
- Can this chain of custody be verified independently, rather than simply asserted?

Neither of the two primary research sources in this report directly measures the state of provenance infrastructure across the AI industry — that is outside their scope. What they establish is the reason this question matters: Shumailov et al. show that the presence or absence of real, verified data is the pivot point between stable and degrading model performance under a replace-style workflow [2]; Kazdan et al. show that the practical value of adding more synthetic data is conditional on how much verified real data is already present [3]. Both results depend on being able to tell real data apart from synthetic data in the first place — which is precisely what provenance tracking is for.

Shumailov et al. flag this directly in their Discussion section, describing it as an open problem: "The need to distinguish data generated by LLMs from other data raises questions about the provenance of content that is crawled from the Internet: it is unclear how content generated by LLMs can be tracked at scale. One option is community-wide coordination to ensure that different parties involved in LLM creation and deployment share the information needed to resolve questions of provenance." [2]

09 Why Trusted Data Could Become More Valuable

It would be an overstatement to say human-generated data is disappearing — the measurement studies in Section 1 describe the composition of *newly published* web content, not a reduction in the total stock of pre-existing human-generated data, most of which remains archived and available. The more precise claim the evidence supports is different: **the total volume of content is very plausibly increasing, while the proportion of that new content with clear, verifiable provenance is not increasing at the same rate** — and may be decreasing, per the measurement studies cited in Section 1.

This distinction is exactly why Shumailov et al. close their paper on a value-of-data note rather than a purely cautionary one: "the value of data collected about genuine human interactions with systems will be increasingly valuable in the presence of LLM-generated content in data crawled from the Internet." [2] Kazdan et al.'s empirical results give this claim a more concrete, testable shape: real data has demonstrable, measurable value that a proportionally larger volume of synthetic data does not substitute for, once real data is already reasonably abundant [3].

KEY FINDING

"Removing synthetic or low-quality data can sometimes bring more value than collecting greater volumes of high-quality data"

Source: Kazdan et al., ICML / PMLR 267, 2025, Section 4 / Discussion — the authors' own conclusion from their real-vs-synthetic data experiments.

Neither paper, however, offers a market-level or pricing analysis of data value — that inference (that verified, trusted data could command a premium as an infrastructure input) is plausible given the technical findings, but it is an extrapolation this report is making explicitly, not a conclusion either source states directly.

10 What Happens Next?

The following scenario analysis separates what the evidence directly shows from what is a reasonable but unproven extrapolation. Each item is labeled by evidence strength, consistent with Figure 6 below.

NEAR TERM

- Continued growth in AI-authored content in newly published web text **OBSERVED**
- Wider industry adoption of AI-content detectors and dataset filtering as a standard pretraining step **PLAUSIBLE**
- Growing practical interest in data-provenance tooling among model developers **PLAUSIBLE**

MEDIUM TERM

- Increasing premium placed on curated, verified-provenance datasets for fine-tuning and pretraining **PLAUSIBLE**
- Stronger data-governance and disclosure norms around training-data sourcing **SPECULATIVE**
- Wider adoption of "accumulate" and fixed-compute-subsample style workflows over full-replace pipelines, since accumulate is shown to avoid collapse across tested settings **SUPPORTED**

LONG TERM

- If real data becomes harder to preserve or identify at scale, recursive feedback effects could reemerge as a genuine risk **PLAUSIBLE**
- An internet in which most new content is AI-generated **OBSERVED trend, unresolved endpoint**
- Inevitable, unavoidable collapse of future AI systems **SPECULATIVE — not established**

CHART 6 — Where the Evidence Is Strong, and Where It Isn't

OBSERVED	AI-detector studies find AI-generated articles are a large and growing share of newly published web text, plateauing near ~50% of sampled Common Crawl articles by late 2025 / early 2026. <i>Source: Graphite/Five Percent (2026)</i>
OBSERVED	Under a replace training-workflow (real data deleted each generation), every task setting tested by both papers (Gaussian modeling, KDE, linear regression, LLM fine-tuning/pretraining) shows measurable performance degradation. <i>Source: Shumailov et al. (2024); Kazdan et al. (2025)</i>
SUPPORTED	Under an accumulate workflow (real data retained, synthetic data added alongside it), test loss on real data stays bounded rather than diverging, across five generative-modeling task settings. <i>Source: Kazdan et al. (2025), Sec. 2-3</i>
SUPPORTED	The value of adding synthetic data depends on how much real data is already available: small amounts of synthetic data can help when real data is scarce, and tend to hurt when real data is already abundant. <i>Source: Kazdan et al. (2025), Sec. 4, Fig. 5</i>
PLAUSIBLE	A fixed compute budget (train on a constant-size subsample of an ever-growing real+synthetic pool) produces slow, bounded degradation — worse than full accumulation but far short of full replace-style collapse. <i>Source: Kazdan et al. (2025), Sec. 3</i>
PLAUSIBLE	As AI-generated content becomes a larger share of newly published web text, verified provenance / “known-human” data may become comparatively more valuable for training — though this is an inference from the above, not a directly measured economic finding. <i>Source: Inference from above sources</i>
SPECULATIVE	Claims that the internet “will become mostly AI-generated,” that AI models “will inevitably collapse,” or that any specific percentage (e.g., 90%) of future web content will be synthetic are not established by the attached evidence. <i>Source: Not supported by attached sources</i>

Figure 6. A summary rubric of the claims made across this report, each graded by the strength of the evidence behind it.

11 Conclusion

The published evidence supports a more specific and more actionable story than “AI is poisoning itself.” Model collapse is real, reproducible, and mathematically well-understood under a specific, narrow condition: recursively training generation after generation of models exclusively on the previous generation's synthetic output, with no real data preserved [2]. Under a more realistic assumption — real data accumulating alongside synthetic data rather than being replaced by it — the same experiments, run by an independent research team across the same and additional task settings, do not show collapse [3]. The determining factor in the evidence is workflow design, not the mere existence of synthetic data.

What the evidence also makes clear is that the value of real data is not static or self-evident — it depends on context, and specifically on how much real data is already available relative to synthetic data being added [3]. That dependency is exactly why the ability to reliably identify, verify, and preserve real, provenance-known data is emerging as a practical concern for anyone building AI training pipelines, independent of whether internet-scale collapse ever occurs.

The future AI bottleneck may not be the ability to generate more data. It may be the ability to know which data is still worth trusting.

This is a research summary, not a forecast. Where the underlying sources report a trend graphically without publishing exact figures, this report has said so rather than estimate a number. Where a claim is plausible but unproven by the attached sources, it has been labeled accordingly. The reader should treat any claim not explicitly tied to a cited source with the same caution this report has tried to apply throughout.

This report was compiled from the primary sources listed in the References section. Two categories of material referenced in the original research brief — a Pew Research dataset and additional unspecified internet AI-content measurement studies — were not available as attached source documents at the time of writing and are therefore not cited; no statistics attributed to those sources appear in this report.

References

- [1] Paredes, J. L., Druck, G., Benson, B., & Smith, E. "AI Now Writes as Many Online Articles as Humans." *Five Percent — Graphite Research*, May 2026. Common Crawl-based study of 55,400 English-language web articles (Jan 2020–Mar 2026), classified using Pangram, Copyleaks, and GPTZero AI detectors. URL: <https://graphite.io/five-percent/research/ai-now-writes-as-many-online-articles-as-humans-do>
- [2] Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. "AI models collapse when trained on recursively generated data." *Nature*, Vol. 631, pp. 755–759, 2024. University of Oxford / University of Cambridge / Imperial College London / University of Toronto / University of Edinburgh. DOI: 10.1038/s41586-024-07566-y. URL: <https://doi.org/10.1038/s41586-024-07566-y>
- [3] Kazdan, J., Schaeffer, R., Dey, A., Gerstgrasser, M., Rafailov, R., Donoho, D., & Koyejo, S. "Collapse or Thrive? Perils and Promises of Synthetic Data in a Self-Generating World." *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, PMLR 267, 2025. Stanford University, Departments of Statistics and Computer Science. URL: <https://openreview.net/forum?id=5B2K4LRgmz> (referenced Gerstgrasser et al. prior work); primary paper as attached source document.
- [4] Originality.ai. "Amount of AI Content in Google Search Results." Ongoing longitudinal study, data points March 2024–September 2025. URL: <https://originality.ai/ai-content-in-google-search-results>
- [5] Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Korbak, T., Sleight, H., Agrawal, R., Hughes, J., Pai, D. B., Gromov, A., Roberts, D., Yang, D., Donoho, D. L., & Koyejo, S. "Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data." 2024. Cited within [3] as the source of Claims 1 and 2 tested therein. URL: <https://openreview.net/forum?id=5B2K4LRgmz>