The Internet Is Eating Itself

Model collapse, synthetic sludge, and the slow poisoning of the data that made AI smart.

Kymata Labs Research · An independent research institution studying how AI systems actually behave in production.

~12 min read

Tags · AI Training Data · Model Collapse · Synthetic Data · Provenance · AI Capability

The best AI models were built on a one-time gift: a vast, messy, mostly human internet, written before anyone thought a machine would read it. That internet is gone. The new one is filling, by the day, with text and images that AI wrote, and when you feed a model its own output, it gets worse. This is not a metaphor. It has a name in the literature, *model collapse*, and the frightening part has nothing to do with machines turning on us. It is quieter than that: the well that made them smart is filling with their own runoff, and they have to keep drinking from it. This paper traces the feedback loop from synthetic data to measurable degradation, sizes how contaminated the commons already is, and argues that verified, point-in-time human data is about to become the scarcest, most defensible input in the entire stack.

The argument, in five moves

1. **The training set was a one-time gift.** Today's frontier models learned from a human internet written before anyone performed for the crawler. That naïveté is exactly what made the data clean.
2. **Feeding a model its own output degrades it.** Indiscriminate, recursive training on machine-generated text causes *model collapse*: the tails of the original distribution disappear, and the model narrows generation after generation.
3. **The open web is drifting toward that exact recipe.** Roughly half of newly published articles are already machine-made, a large share of the multilingual web is machine-translated, and reliable AI-text detection at web scale does not exist.
4. **Clean human data is now priced and fought over.** When a resource is abundant, nobody licenses it. Verified human text is now the subject of nine-figure deals and user revolts, which tells you it has become scarce.
5. **Provenance is the defensible asset.** The studios that can hoard verified, human, point-in-time data, and keep their corpora clean, compound an advantage that a collapsing web cannot manufacture and a competitor cannot crawl their way back into.

The one-liner: the danger isn't a rogue intelligence. It's that we may be living through the smartest these systems will ever be, because the input that produced them is finite and contaminating.

A result, on the cover of Nature

In July 2024, *Nature* ran a study as its cover story. The finding, in the authors' own words: "indiscriminate use of model-generated content in training causes irreversible defects in the resulting models, in which tails

of the original content distribution disappear," an effect they named *model collapse* [1]. They demonstrated it across large language models, variational autoencoders, and Gaussian mixture models, three very different architectures, the same failure.

The canonical illustration is brutal in its simplicity. An OPT-125m model, retrained over and over on its own writing, began Generation 0 with coherent prose about medieval architecture and degraded by Generation 9 into repetitive nonsense about multicoloured "jackrabbits." The rare, the specific, the long-tail, the parts of reality that make a model accurate rather than merely plausible, washed out first. Degradation is invisible from inside any single generation: each model looks competent on its own benchmarks. It is only across generations, on the long tail of human expression, that the narrowing becomes legible.

Replacement, not synthetic data as such

Read the result carefully, because the distinction is the whole argument. The *Nature* collapse is about a specific, dangerous recipe: indiscriminate, **replacement** training, where each generation learns chiefly from the previous one's output. A companion line of work sharpens the warning rather than softening it. Gerstgrasser and colleagues showed that when synthetic data **accumulates alongside** the original human data instead of replacing it, collapse is bounded, because the human floor holds [2].

So synthetic data is not poison on its own. The poison is the trajectory: a web where verified human text thins out, model output thickens, and training proceeds indiscriminately because nobody can reliably tell the two apart. That is precisely the trajectory the open web is on.

The measured web: about half of new articles are machine-made

How close is that trajectory? Closer than comfortable. Graphite's analysis found that AI-generated articles reached rough parity with human-written ones by late 2024, then plateaued near a fifty-fifty split through Q1 2026 [3]. Be precise about the claim: this is *roughly half of newly published web articles*, not "the majority of the web," and not the whole internet. But it is the freshest stratum, the layer a crawler scrapes next.

Contamination is not only fresh AI text. AWS AI Labs researchers built a multi-way parallel corpus from the web and found that 57.1% of sentences sit in multi-way-parallel (mass machine-translated) clusters, the same low-quality content pumped through translation into many languages at once [4]. Low-quality material is disproportionately machine-translated, skewing the multilingual web toward exactly the garbled, hallucination-prone text you least want in a training set. The poison was already in the groundwater before the chatbots arrived; the chatbots are raising the level.

Signal	Figure	Source
Newly published web articles that are AI-generated	~50%	Graphite (2024–2026)
Web sentences in mass machine-translated clusters	57.1%	Thompson et al., ACL 2024
High-quality human text stock fully utilized	~2028	Epoch AI projection

Figures are stated as published. The ~50% is a share of newly published articles, not the web overall.

Clean human data now has a price

Watch where the money goes and you learn what is actually scarce. Verified human content is now priced and fought over. Reddit licensed its human conversation to Google for a reported ~$60 million a year [5]. OpenAI signed News Corp for a reported more than $250 million over five years [6]. And in the most telling episode of all, when Stack Overflow struck its OpenAI deal in 2024, some users sabotaged or deleted their own answers in protest, a small and furious referendum on who the machines were trained on. You do not pay nine figures, and users do not revolt, over an abundant resource.

Deal	Reported value	When
Reddit ⟶ Google (data licensing)	~$60M / year	Feb 2024
News Corp ⟶ OpenAI (content)	>$250M / 5 years	May 2024
Stack Overflow ⟶ OpenAI (+ user protest)	undisclosed	May 2024

“Peak data” arrives around 2028

Epoch AI put a date on the scarcity. Their updated projection estimates the stock of high-quality public human text will be fully utilized somewhere between 2026 and 2032, central estimate around 2028 [7]. Their own hedge matters, and we keep it: they place only about a 20% chance that scaling slows significantly by 2040, because synthetic data done carefully, better data efficiency, and new modalities may yet relax the limit. Read together with model collapse, the forecast is sobering rather than apocalyptic: the human well runs low right as the synthetic runoff runs high.

It is not just text: the images go MAD too

Lest this seem a quirk of language, the same disease shows up in pictures. Researchers at Rice and Stanford coined **MAD**, for Model Autophagy Disorder: a self-consuming loop in which image generators trained on their own (and each other’s) outputs lose quality and diversity, generation after generation, unless a sufficient stream of fresh real images keeps flowing in [8]. Autophagy is a system eating itself. The name is not subtle, and it isn’t meant to be.

The gift was free. That was the whole problem.

Today’s best models were trained on a corpus no one will ever assemble again: decades of human writing produced **before** the writers knew a machine would harvest it. Forum arguments, Wikipedia edits, recipe blogs, code comments, half-finished novels, billions of people writing for each other with no thought of training data. That naïveté is what made the data clean. It was a snapshot of human language taken while no one was performing for the camera.

Then the cameras turned on. The moment models could write fluently, the cheapest way to fill a web page became a prompt, and because that text is free, fast, and good enough, it floods in, onto the same open web the next model will scrape. The supply chain quietly closed into a loop: the model’s output becomes the model’s input. The internet stopped being a record of human thought and started becoming a mirror the machines hold up to themselves. Nothing about this was a decision. It is an emergent property of cheap

generation meeting an open commons, a tragedy of the commons where the grass is verified human data and everyone, including the machines, is grazing.

Two kinds of model-builders are forming

Model collapse does not fall evenly. It sorts the field into two camps. On one side, the labs that **own or can buy verified human data**, whether the Reddit license, the News Corp archive, or a proprietary stream of real human interaction, paired with the discipline to keep synthetic data **accumulating alongside** the real, never replacing it. They keep the human floor under their models, and the floor holds.

On the other side, everyone scraping the open web in good faith, training **indiscriminately** on a corpus they can no longer clean, because reliable AI-text detection at web scale doesn't exist and gets harder as the models improve. The ambition is identical; the trajectory runs the opposite way. The first group's models stay sharp. The second group's models slowly forget the tails, and, fluent to the last, won't show the damage until something specific and true is asked of them and the answer comes back confident and wrong. The scarce resource was never compute. It is clean water.

What it means, read by three different readers

Collapse is not destiny. The same research that diagnoses it also names the cure: keep real human data in the mix, and keep the corpus clean. What that demands depends on who you are.

For individuals
Treat fluent output as a draft, not a verdict, especially on the long-tail and the specific. The failure mode is confidence, not gibberish.

For employers and builders
Audit your data provenance the way you audit your dependencies. If you train or fine-tune, the question is no longer "how much data" but "how much **verified human** data, and can you prove the corpus is clean?" Provenance is becoming a line item.

For policymakers
The asset worth protecting is the human web itself, and the incentives that keep people writing it. Provenance, labeling, and the economics of verified data are infrastructure questions, not content-moderation footnotes.

Frequently asked questions

Wait, didn't a 2024 study show synthetic data is fine, even helpful?
Both things are true, and the distinction is the whole paper. The *Nature* result is about indiscriminate, replacement training: each generation learns mostly from the last one's output, and the model collapses. A separate study (Gerstgrasser et al.) showed that when synthetic data accumulates alongside the original human data rather than replacing it, collapse is bounded, because the human floor holds. So synthetic data isn't poison by itself. The danger is in the recipe the open web is drifting toward: less and less verified human data, more and more model output, trained on indiscriminately because nobody can tell which is which anymore.

Is the whole internet really written by AI now?
No, and we're careful with that number. Graphite's analysis found that roughly half of newly published web articles are AI-generated, reaching rough parity in late 2024 and then plateauing near 50/50. That's newly published articles, not the web as a whole, and not "the majority." The point isn't that humans stopped

writing. It's that the freshest layer of the web, the part a crawler scrapes tomorrow, is now about half machine-made, and rising AI content is harder to label and filter than it looks.

Can't they just filter the AI-generated text back out before training?

In theory. In practice, reliable detection of AI text at web scale is an unsolved problem, and it gets harder as models get better, since the whole point of a good model is that its output is indistinguishable from human writing. Worse, contamination hides where filters don't look: machine-translated pages dressed up as native content, AI text lightly edited by a human, synthetic data laundered through a dozen reposts. You can't cleanly remove what you can't reliably detect.

Does this mean AI progress is about to stop?

Unsure, and we won't pretend otherwise. Epoch AI, which projects the high-quality human text stock is fully utilized between 2026 and 2032 (central estimate ~2028), still puts only about a 20% chance on scaling significantly slowing by 2040. Synthetic data done carefully, better data efficiency, and new modalities may relax the constraint. The honest claim is narrower: the cheap, clean, human-made data that produced today's best models is finite and contaminating, and the strategies that work around that are more expensive and more fragile than the free buffet that got us here.

What's the single most important idea to take away?

Clean human data is now the scarce input, and it's being diluted by the very systems that depend on it. The models didn't run out of compute or ideas. What they're running short of is uncontaminated water at the well. Whoever controls verified human data, and whoever can keep training corpora clean, holds the real leverage in the next phase of AI.

References

1 Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R. & Gal, Y. (2024). "AI models collapse when trained on recursively generated data." *Nature*, 631, 755–759. (Cover story, 25 July 2024.) https://www.nature.com/articles/s41586-024-07566-y

2 Gerstgrasser, M. et al. (2024). "Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data." arXiv:2404.01413. https://arxiv.org/abs/2404.01413

3 Graphite (2025). "More Articles Are Now Created by AI Than Humans," with follow-up analysis showing AI- and human-written share plateaued near parity through Q1 2026. https://graphite.io/five-percent/more-articles-are-now-created-by-ai-than-humans

4 Thompson, B., Dhaliwal, M. P., Frisch, P., Domhan, T. & Federico, M. (2024). "A Shocking Amount of the Web Is Machine Translated: Insights from Multi-Way Parallelism." Findings of the ACL 2024. https://aclanthology.org/2024.findings-acl.103/

5 Reuters (22 Feb 2024). "Exclusive: Reddit in AI content licensing deal with Google." (Reported ~$60M / year.) https://www.reuters.com/technology/reddit-ai-content-licensing-deal-with-google-sources-say-2024-02-22/

6 The Wall Street Journal (22 May 2024). "News Corp and OpenAI Sign Landmark Multi-Year Global Partnership." (Reported >$250M / 5 years.) https://www.wsj.com/business/media/openai-news-corp-strike-deal-23f186ba

7 Villalobos, P. et al. / Epoch AI (2022, updated June 2024). "Will we run out of data? Limits of LLM scaling based on human-generated data." arXiv:2211.04325. https://epoch.ai/blog/will-we-run-out-of-ml-data-evidence-from-projecting-dataset

8 Alemohammad, S. et al. (2024). "Self-Consuming Generative Models Go MAD." International Conference on Learning Representations (ICLR) 2024. https://arxiv.org/abs/2307.01850