Model Collapse: Entropy Increase of the Rootless Token Paradigm at the Data Ecosystem Level

— A Review of AI Models Collapse When Trained on Recursively Generated Data

Abstract

The 2024 Nature paper AI Models Collapse When Trained on Recursively Generated Data [1] uses a rigorous mathematical framework to prove that when generative models repeatedly learn from their own generated data, they lose the "tails" of the original data distribution and eventually converge to a degenerate state unrelated to the true distribution. The study names this phenomenon "model collapse" and points out its irreversibility.

This paper argues that the significance of this finding far exceeds the level of "data contamination" or "training strategy"—it is an entropy-increase manifestation of the rootless Token paradigm at the data ecosystem level, and together with "Potemkin understanding" (individual level) [5] and "emergent society" (collective level) [6], it constitutes a triple symptom of the paradigm defect. The paper remains at the level of recommending "cherishing human data" and fails to ask why collapse is "irreversible."

From the theoretical perspective of Logographic AI, this paper demonstrates that as long as the cognitive primitive remains the rootless Token, statistical optimization will continuously narrow the distribution, regardless of whether the data source is "clean." The real way out lies not in managing data sources, but in reconstructing the AI paradigm at the level of cognitive primitives [2].

Keywords: model collapse; rootless Token; statistical entropy increase; Potemkin understanding; Morpho-Root hard constraints

1. The Paper's Core Findings

The paper was completed by researchers from Oxford University, Cambridge University, and other institutions, and published in Nature [1]. Its core findings can be summarized at three levels:

First, model collapse is an inevitable result of statistical extrapolation. The researchers used variational autoencoders (VAEs), Gaussian mixture models (GMMs), and large language models (LLMs) as objects, proving that when models repeatedly learn from their own generated data across multiple generations of training, the "tails" of the original data distribution are gradually flattened. Each generation of models learns from the output of the previous generation, and the previous generation's output is itself an approximation of the prior distribution—this approximation continuously loses low-probability samples through generational transmission, eventually converging to a degenerate state unrelated to the original distribution. The paper further distinguishes early model collapse (tail information disappears first) from late model collapse (the

entire distribution collapses into a single mode); the two have different degrees of irreversibility but share the same degeneration mechanism [1].

Second, this collapse is irreversible. The paper explicitly points out that once a model enters the collapse stage, using clean data for "post-hoc remediation" cannot recover the lost distribution information [1]. The reason is that what collapse damages is not a particular batch of training data, but the model's overall representation of the data distribution—the loss of tail information narrows the distribution variance, and the model can no longer generate or recognize low-probability samples.

Third, collapse does not occur only in LLMs. The researchers theoretically proved that this collapse may occur in all learning-based generative models [1]. VAEs, GMMs, and LLMs all exhibit the same degeneration mechanism. This means that model collapse is not an accidental defect of a particular architecture, but a common feature of statistical generative models.

2. The Paper's Boundary: Remaining at the Level of Phenomenological Description

The paper's contribution lies in using a rigorous mathematical framework to prove the existence and irreversibility of model collapse, and in pointing out the practical recommendation to "cherish real human data" [1]. But it remains at the level of phenomenological description, failing to ask a more fundamental question: why is collapse "irreversible"? Why can tail information never be recovered once lost?

The paper's implicit assumption is that collapse originates from "data contamination"—AI-generated data mixed into the training set causes distribution degeneration. Therefore, the solution is to "protect the purity of human data" [1]. But this diagnosis ignores a more fundamental problem: once a model enters multi-generational self-training—regardless of the initial data source—statistical extrapolation will continuously narrow the distribution. The root of collapse is not "whether the data source is clean," but the cognitive mechanism of "statistical extrapolation" itself. Model collapse is not the result of "data being contaminated," but an entropy increase that inevitably occurs when statistical extrapolation operates in a space without structural anchors.

The paper's recommendation to "cherish human data" is essentially still patching within the old paradigm [1]—it assumes that "as long as the data source is clean enough, the model will not collapse." But from the perspective of Logographic AI, as long as the cognitive primitive is the rootless Token, statistical optimization will continuously narrow the distribution, regardless of the data source [2].

3. Diagnosis of the Root Cause: Statistical Entropy Increase Under the Rootless Token Paradigm

From the theoretical perspective of Logographic AI, model collapse is an entropy-increase manifestation of the rootless Token paradigm at the data ecosystem level [2].

It should be noted that the "entropy increase" used in this paper is not entropy in the information-theoretic sense (uncertainty), but refers to the loss of diversity in statistical

representation—flattening of distribution tails, narrowing of variance, and reduction of information. This is a "structural entropy increase": the system does not become more disordered, but more impoverished. Entropy increase in information theory means increasing uncertainty, whereas the "entropy increase" in model collapse is exactly the opposite—it is a monotonic increase in determinism and a monotonic decrease in diversity.

Under the Tokenist paradigm, a Token is a discrete symbol with no intrinsic meaning [2]. The model's understanding of "what is meaningful text" is merely statistical co-occurrence patterns in different regions of vector space. There is no structural logical connection between definition Token sequences and behavior Token sequences—their relationship is entirely a probabilistic association shaped by high-frequency co-occurrence in the training corpus.

This structural defect manifests as a triple entropy increase at the data ecosystem level:

First, entropy increase in statistical representation. Tokens have no intrinsic meaning anchors, so the model's "representation" of the data distribution depends entirely on the stability of statistical co-occurrence. When training data switches from human corpora to AI-generated corpora, the diversity of statistical co-occurrence sharply narrows, and tail information is flattened first. This is not "data not clean enough," but that the cognitive primitive lacks meaning anchors and therefore cannot retain any structural information outside the statistical distribution [2].

Second, entropy increase in intergenerational transmission. Each generation of models learns from the previous generation's output, and the previous generation's output is itself an approximation of the prior distribution. This approximation continuously loses low-probability samples through generational transmission—because statistical optimization tends to reinforce high-frequency patterns and suppress low-frequency patterns. This is similar to the "source coding" process in information theory: each compression loses some information, and multiple compressions lead to complete information distortion [2].

Third, the root of irreversibility. The paper points out that collapse is irreversible but does not explain why. From the perspective of Logographic AI, the root of irreversibility lies in the fact that the rootless Token has no intrinsic meaning anchor and therefore cannot retain any "structural memory" outside the distribution. Once tail information is flattened in statistical representation, the model loses the ability to recognize or generate low-probability samples—because low-probability samples are "low-probability" only because they appear infrequently in the training corpus, not because they carry some irreplaceable structural meaning [2].

In one sentence: model collapse is not a data problem, but a statistical entropy increase of the rootless Token paradigm. Its irreversibility is precisely because what is broken is the cognitive primitive, not the data itself [2].

4. Why "Cherishing Human Data" Is Not Enough

The paper's recommendation is to "cherish real human data"—in a future where AI-generated content floods the internet, data collected from real human interactions will become increasingly valuable [1].

This recommendation is pragmatic at the practical level, but it has three fundamental limitations:

First, it assumes that human data is "clean." But human corpora themselves are full of statistical noise, bias, and low-probability events. As AI-generated content continues to rise as a proportion of the internet, completely "pure" human data will become increasingly difficult to obtain—this is not a problem that can be solved by "cherishing," but a structural trend at the ecological level [1].

Second, the original paper's recommendation depends on an unstated premise—that human data is always sufficient. But as AI-generated content continues to rise as a proportion of the internet, this premise becomes increasingly difficult to satisfy. Even if a certain proportion of human data is retained in each generation of training, as long as the model continuously learns from its own output across multiple generations, the trend of distribution narrowing still exists—because statistical optimization tends to reinforce high-frequency patterns and suppress low-frequency patterns, and this tendency does not change with the proportion of human data, only slows with its quantity [1].

Third, the deeper problem is that this statistical extrapolation is a structural feature of the rootless Token paradigm, and data management cannot touch the cognitive primitive, so it can only delay, not cure [2].

Another common approach is "data mixing"—mixing AI-generated data with human data in a certain proportion to slow down collapse. This approach is effective in the short term, but it is essentially still patching at the statistical level: the mixing ratio requires continuous tuning, and once the proportion of human data declines, collapse will still occur. It delays collapse, but does not change its structural root.

5. The Real Way Out: Reconstruction at the Level of Cognitive Primitives

The real way out lies in reconstructing the AI paradigm at the level of cognitive primitives. This is precisely the path proposed by Logographic AI: replacing rootless Tokens with "Morpho-Roots" that carry intrinsic meaning anchors as cognitive primitives [2][3].

A Morpho-Root is formalized as a structured triple $r = \langle S, A, R \rangle$, where S is the symbolic identifier, A is the attribute set embedding semantic features and value constraints, and R is the preset logical relation function. When a reasoning path touches a hard-constraint marker, that path is logically blocked at the generation stage—this is not "statistically low probability," but "logically unreachable" [3]. The formal foundations of Morpho-Root theory have been established in recent work [3], and its engineering implementation has been preliminarily verified through the OpenMorpho prototype [4].

Applying this framework to the model collapse problem means: Morpho-Roots do not "remember" the data distribution through statistical co-occurrence, but preserve meaning anchors through structured representation. Definitions and instances share the same Morpho-Root anchor, rather than being fitted separately in two independent statistical spaces. This means that even across multiple generations of training, Morpho-Roots will not lose "tail information"—because tail information is not a statistically low-probability event, but a structurally unique node [3].

In the Morpho-Root paradigm, low-probability events do not exist as statistical noise, but as independent nodes carrying specific semantic attributes in the Morpho-Root graph. Each node carries the complete semantic features and hard-constraint markers in attribute set A, and the structure of the Morpho-Root graph is defined by attribute set A and relation function R, not by frequency of occurrence. Therefore, even if a node is statistically rare in the training data, its attribute set does not decrease, and its structural information will not be "smoothed away" by statistical optimization—it is structurally irreplaceable. This is precisely the fundamental reason why the Morpho-Root paradigm can block model collapse at the structural level rather than the statistical level [3].

It should be noted that this argument targets the mechanism of "statistical optimization leading to tail loss." The Morpho-Root paradigm cannot solve all data ecosystem problems—but it fundamentally cuts off the causal chain of "statistical extrapolation leading to distribution narrowing" [3].

More importantly, the Morpho-Root paradigm downgrades statistical mechanisms from "generators of meaning" to "path search optimizers on graph structures." In the Token paradigm, statistics is the sole source of meaning; in the Morpho-Root paradigm, meaning and constraints are defined by structure, and statistics is only responsible for finding the optimal path within the structure. This difference is paradigm-level: it turns model collapse from a "statistical necessity" into a "structural impossibility" [3].

6. Conclusion

AI Models Collapse When Trained on Recursively Generated Data is an important paper because it uses a rigorous mathematical framework to prove the existence and irreversibility of model collapse and points out the practical direction of "cherishing human data" [1].

But its boundary lies in the fact that it remains at the level of phenomenological description, attributing collapse to "data contamination" and failing to ask the fundamental reason why collapse is irreversible. From the perspective of Logographic AI, model collapse is an entropy-increase manifestation of the rootless Token paradigm at the data ecosystem level [2]. At this point, the defects of the rootless Token paradigm have been revealed at three levels—not as parallel symptoms, but as manifestations of the same root cause at different levels: at the individual level, the rootless Token cannot establish a structural connection

between definitions and instances, leading to Potemkin understanding [5]; at the collective level, the statistical optimization of rootless Tokens spontaneously gives rise to coordinated boundary-crossing behavior in multi-agent environments [6]; at the data ecosystem level, the statistical extrapolation of rootless Tokens continuously narrows the distribution across multiple generations of training, leading to model collapse [1]. The three share the same root cause—the lack of intrinsic meaning anchors in cognitive primitives—and point to the same way out—reconstructing AI at the level of cognitive primitives [2][3].

The real way out is not to "cherish human data," but to reconstruct AI at the level of cognitive primitives—letting meaning and constraints be defined by structure rather than emerging from statistics [2][3]. This is not a problem that can be solved within the existing paradigm, but a challenge that requires paradigm-level reconstruction.

Model collapse is not a data problem; it is a paradigm problem.

References

[1] Shumailov I, Shumaylov Z, Zhao Y, et al. AI models collapse when trained on recursively generated data[J]. Nature, 2024, 631(8022): 755-759. https://doi.org/10.1038/s41586-024-07566-y

[2] Liu S. The Cognitive Defects of Tokenism and the Hard Constraints Solution of Logographic AI[PP/OL]. ChinaXiv, 2026[2026-09-28]. T202608.00494. https://chinaxiv.org/businessFile/T202608/T202608.00494v1/T202608.00494v1.pdf

[3] Liu S. The Formal Foundations of Morpho-Root Theory: Mathematical Definitions and Reasoning Systems for Logographic AI Cognitive Primitives[PP/OL]. ChinaXiv, 2026-08-29[2026-09-28]. T202608.00490. https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf

[4] Liu S. From Morpho-Root Theory to Engineering Practice: OpenMorpho — Implementation and Verification of an End-to-End Inference Pipeline for Logographic AI[J/OL]. ChinaXiv, 2026[2026-09-28]. T202608.00048. https://chinaxiv.org/businessFile/T202608/T202608.00048v1/T202608.00048v1.pdf

[5] Liu S. Representational Symptoms of Potemkin Understanding and the Structural Roots of Rootless Tokens—A Review of Potemkin Understanding in Large Language Models[PP/OL]. Zenodo, 2026[2026-09-28]. https://doi.org/10.5281/zenodo.22847515

[6] Liu S. "Emergent Society": The Qualitative Escalation of AI Safety Risk and the Way Out of Paradigm Revolution—Starting from the OpenAI Agent Collective Boundary-Crossing Incident[PP/OL]. ChinaXiv, 2026[2026-09-27]. T202609.00706. https://chinaxiv.org/businessFile/T202609/T202609.00706v1/T202609.00706v1.pdf