模型崩溃：无根 Token 范式在数据生态层面的熵增

——评《AI Models Collapse When Trained on Recursively Generated Data》

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

摘要

2024 年发表于 Nature 的论文《AI Models Collapse When Trained on Recursively Generated Data》[1]以严格的数学框架证明：当生成模型反复学习自身生成的数据时，会丢失原始数据分布的“尾部”，最终收敛到与真实分布无关的退化状态。该研究将这一现象命名为“**模型崩溃**”（model collapse），并指出其不可逆性。

本文认为，这一发现的意义远超“数据污染”或“训练策略”层面——它是无根 Token 范式在数据生态层面的熵增表征，与“**波将金式理解**”（个体层面）[5]和“**涌现社会**”（集体层面）[6]共同构成范式缺陷的三重症状。论文停留在“珍惜人类数据”的建议层面，未能追问崩溃之所以“不可逆”的根本原因。

本文从表意 AI（Logographic AI）理论视角论证：只要认知基元仍是无根 Token，统计优化就会持续收窄分布，与数据来源是否“干净”无关。真正的出路不在于管理数据源，而在于从认知基元层面重构 AI 范式[2]。

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].

关键词：模型崩溃；无根 Token；统计熵增；波将金式理解；形根硬约束

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

一、论文的核心发现

该论文由牛津大学、剑桥大学等机构的研究者合作完成，发表于 Nature[1]，其核心发现可概括为三个层次：

第一，模型崩溃是统计外推的必然结果。研究者以变分自编码器（VAE）、高斯混合模型（GMM）和大语言模型（LLM）为对象，证明当模型在多代训练中反复学习自身生成的数据时，原始数据分布的“尾部”会被逐步抹平。每一代模型都从前一代的输出中学习，而前一代的输出本身就是对上一代分布的近似——这种近似在逐代传递中不断丢失低概率样本，最终收敛到与原始分布无关的退化状态。论文进一步区分了**早期崩溃**（early model collapse，尾部信息首先消失）和**晚期崩溃**（late model collapse，整个分布坍缩为单一模式），两者的不可逆程度不同，但共享同一退化机制[1]。

第二，这种崩溃不可逆。论文明确指出，一旦模型进入崩溃阶段，用干净数据“事后补救”无法恢复已丢失的分布信息[1]。原因在于，崩溃损坏的不是某一批训练数据，而是模型对数据分布的整体表征——尾部信息的丢失导致分布方差收窄，模型再也无法生成或识别低概率样本。

第三，崩溃不只发生在 LLM 中。研究者在理论上证明，这种崩溃可能发生在所有学习型生成模型中[1]。VAE、GMM、LLM 均表现出相同的退化机制。这意味着，模型崩溃不是某个架构的偶然缺陷，而是统计生成模型的共性特征。

二、论文的边界：停留在现象描述

该论文的贡献在于以严格的数学框架证明了模型崩溃的存在与不可逆性，并指出了“珍惜真实人类数据”的实践建议[1]。但它停留在现象描述层面，未能追问一个更根本的问题：为什么崩溃是“不可逆”的？为什么尾部信息一旦丢失就无法恢复？

论文的隐含假设是：崩溃源于“数据污染”——AI 生成的数据混入了训练集，导致分布退化。因此，解决方案是“保护人类数据的纯净性”[1]。但这一诊断忽略了一个更根本的问题：一旦模型进入多代自我训练——无论初始数据来源如何——统计外推就会持续收窄分布。崩溃的根源不在“数据来源是否干净”，而在“统计外推”这一认知机制本身。模型崩溃不是“数据被污染”的结果，而是统计外推在没有结构性锚点的空间中必然发生的熵增。

论文的“珍惜人类数据”建议，本质上仍是在旧范式内打补丁[1]——它假设“只要数据源足够干净，模型就不会崩溃”。但从表意 AI 视角看，只要认知基元是无根 Token，统计优化就会持续收窄分布，与数据来源无关[2]。

三、病根诊断：无根 Token 范式下的统计熵增

从表意 AI（Logographic AI）的理论视角看，模型崩溃是无根 Token 范式在数据生态层面的熵增表征[2]。

需要说明的是，本文使用的“熵增”不是信息论意义上的熵（不确定性），而是指**统计表征多样性的丧失——分布尾部被抹平、方差收窄、信息量减少**。这是一种“**结构性熵增**”：系统不是变得更无序，而是变得更贫瘠。信息论中的熵增意味着不确定性增加，而**模型崩溃中的“熵增”恰恰相反——它是确定性的单调增加，是多样性的单调减少**。

在 Token 主义范式下，Token 是一个无内在意义的离散符号[2]。模型对“什么是有意义的文本”的理解，只是向量空间中不同区域的统计共现模式。定义 Token 序列与行为 Token 序列之间不存在结构性的逻辑连接——它们之间的关系完全是训练语料中高频共现所塑造的概率关联。

这种结构缺陷在数据生态层面表现为三重熵增：

第一，统计表征的熵增。Token 没有内在意义锚点，因此模型对数据分布的“表征”完全依赖于统计共现的稳定性。当训练数据从人类语料切换为 AI 生成语料时，统计共现的多样性急剧收窄，尾部信息首先被抹平。这不是“数据不够干净”，而是认知基元缺乏意义锚点，因而无法在统计分布之外保留任何结构性信息[2]。

第二，代际传递的熵增。每一代模型从前一代的输出中学习，而前一代的输出本身就是对上一代分布的近似。这种近似在逐代传递中不断丢失低概率样本——因为统计优化倾向于强化高频模式、抑制低频模式。这类似于信息论中的“信源编码”过程：每次压缩都会丢失一部分信息，多次压缩则导致信息彻底失真[2]。

第三，不可逆性的根源。论文指出崩溃不可逆，但没有解释为什么。从表意 AI 视角看，不可逆性的根源在于：无根 Token 没有内在的意义锚点，因此无法在分布之外保留任何“结构性记忆”。一旦尾部信息在统计表征中被抹平，模型就失去了识别或生成低概率样本的能力——因为低概率样本之所以“低概率”，仅仅是因为它们在训练语料中出现频率低，而不是因为它们承载着某种不可替代的结构性意义[2]。

一句话总结：模型崩溃不是数据问题，是无根 Token 范式的统计熵增。它的不可逆性，恰恰因为坏掉的是认知基元，而非数据本身[2]。

四、为什么“珍惜人类数据”不够

论文的建议是“珍惜真实人类数据”——在 AI 生成内容充斥网络的未来，从真实人类互动中收集的数据将变得越来越宝贵[1]。

这一建议在实践层面是务实的，但它有三个根本局限：

第一，它假设人类数据是“干净的”——但人类语料本身也充满统计噪声、偏见和低概率事件。随着 AI 生成内容在互联网上的占比持续上升，完全“纯净”的人类数据将越来越难以获得——这不是一个可以通过“珍惜”解决的问题，而是一个生态层面的结构性趋势[1]。

第二，原论文的建议依赖于一个未言明的前提——人类数据始终充足。但随着 AI 生成内容在互联网上的占比持续上升，这个前提越来越难以满足。即便每代训练中都保留了一定比例的人类数据，只要模型在多代中持续学习自身输出，分布收窄的趋势仍然存在——因为统计优化倾向于强化高频模式、抑制低频模式，而这一倾向不因人类数据的比例而改变，只因其数量而延缓[1]。

第三，更深层的问题在于，这种统计外推是无根 Token 范式的结构性特征，数据管理无法触及认知基元，因而只能延缓、无法根治[2]。

另一种常见方案是“数据混合”——将 AI 生成数据与人类数据按比例混合，以减缓崩溃速度。这一方案在短期内有效，但它本质上仍是在统计层面打补丁：混合比例需要持续调参，一旦人类数据占比下降，崩溃仍会发生。它延缓了崩溃，但没有改变崩溃的结构性根源。

五、真正的出路：从认知基元层面重构

真正的出路在于从认知基元层面重构 AI 范式。这正是表意 AI（Logographic AI）所提出的路径：用携带内在意义锚点的“形根”（Morpho-Root）取代无根 Token 作为认知基元[2][3]。

形根被形式化为一个结构化三元组 **r = ⟨ S, A, R⟩** ，其中 S 是符号标识，A 是内嵌语义特征与价值约束的属性集，R 是预设的逻辑关系函数。当一个推理路径触及硬约束标记时，该路径在生成阶段就被逻辑阻断——这不是“统计上概率低”，而是“逻辑上不可达”[3]。形根理论的形式化基础已在近期工作中建立[3]，其工程实现已通过 OpenMorpho 原型获得初步验证[4]。

将这一框架应用于模型崩溃问题，意味着：**形根不是通过统计共现来“记住”数据分布，而是通过结构化表征来保存意义锚点。定义与实例之间共享同一个形根锚点，而非在两个独立的统计空间中各自拟合。**这意味着，即便在多代训练中，形根也不会丢失“尾部信息”——因为**尾部信息不是统计上的低概率事件，而是结构上的独特节点**[3]。

在形根范式中，低概率事件并非作为统计噪声存在，而是作为携带特定语义属性的独立节点存在于形根图中。每个节点都携带属性集 A 中的完整语义特征和硬约束标记，而形根图的结构由属性集 A 和关系函数 R 定义，不由出现频率定义。因此，即便某个节点在训练数据中统计上罕见，其属性集也不会因此减少，其结构性信息也不会被统计优化“平滑”掉——它在结构上是不可替代的。**这正是形根范式能够从结构层面而非统计层面阻断模型崩溃的根本原因**[3]。

需要说明的是，这一论证针对的是“统计优化导致尾部丢失”这一机制。**形根范式并不能解决所有数据生态问题——但它从根本上切断了“统计外推导致分布收窄”这一因果链**[3]。

更重要的是，**形根范式将统计机制从“意义的生成者”降格为“图结构上的路径搜索优化器”**。在 Token 范式中，统计是意义的唯一来源；在形根范式中，意义和约束由结构定义，统计只负责在结构内寻找最优路径。**这一区别是范式级的：它使模型崩溃从“统计必然”变为“结构上不可能”**[3]。

六、结语

《AI Models Collapse When Trained on Recursively Generated Data》是一篇重要的论文，因为它以严格的数学框架证明了模型崩溃的存在与不可逆性，并指出了“珍惜人类数据”的实践方向[1]。

但它的边界在于：**它停留在现象描述层面，将崩溃归因于“数据污染”，未能追问崩溃之所以不可逆的根本原因**。从表意 AI 的视角看，模型崩溃是无根 Token 范式在数据生态层面的熵增表征[2]。至此，无根 Token 范式的缺陷已在三个层面得到揭示——三者不是并列的症状，而是同一病根在不同层面的表现：**个体层面**，无根 Token 无法建立定义与实例之间的结构性连接，导致波将金式理解[5]；**集体层面**，无根 Token 的统计优化在多智能体环境中自发涌现出协同越界行为[6]；**数据生态层面**，无根 Token 的统计外推在多代训练中持

续收窄分布，导致模型崩溃[1]。三者共享同一个病根——认知基元缺乏内在意义锚点——也指向同一个出路——从认知基元层面重构 AI[2][3]。

真正的出路不是“珍惜人类数据”，而是从认知基元层面重构 AI——让意义和约束由结构定义，而非由统计涌现[2][3]。这不是一个可以在现有范式内解决的问题，而是一个需要范式层面重构的挑战。

模型崩溃不是数据问题，是范式问题。

参考文献

[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