235. THE PROBLEM OF MODEL COLLAPSE: WHY RECURSIVE TRAINING MIGHT LEAD TO THE DECLINE OF LARGE LANGUAGE MODELS

Herasimovich I. I.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Bulavskaya T. V. – Senior Lecturer

The paper examines the phenomenon of model collapse in Large Language Models caused by recursive training on synthetic data. It analyses the statistical mechanics of distributional decay, identifies the progressive stages of knowledge degradation, evaluates the sociological impact on digital platforms, and proposes technical mitigation strategies to preserve the integrity of the artificial intelligence ecosystem.

The artificial intelligence (AI) landscape has reached a critical inflection point where the volume of synthetic, machine-generated content on the open web has officially surpassed original human-authored data. This monumental shift, driven by the cost-efficiency and rapid deployment of modern generative models, has fundamentally altered the digital ecosystem. However, this structural transformation has triggered a systemic, highly detrimental phenomenon known as model collapse. This represents a progressive degenerative process where large language models (LLMs) — advanced artificial intelligence systems trained on massive volumes of internet text to understand and generate human-like language — experience an accelerating degradation in factual reliability and semantic reasoning [1]. As automated web scrapers continually scour the internet to compile training datasets for next-generation algorithms, they inevitably ingest the synthetic outputs produced by previous iterations of AI. Consequently, the digital information ecosystem functions as an Ouroboros — the mythical serpent consuming its own tail. In this self-referential loop, the foundational empirical assumptions of the scaling laws, which previously dictated that simply adding more computational power and data would proportionally increase model intelligence, are being fundamentally challenged.

From a theoretical standpoint, model collapse arises from an inescapable entropy spiral inherent in recursive training loops. Modern autoregressive language models are designed with a mathematically defined objective: to predict the most probable next token in a given sequence. This mechanism naturally favours statistically dominant, highly repetitive linguistic patterns while penalising any deviation from the mathematical norm. Consequently, over successive training generations, these models begin a process of active statistical forgetting, targeting the tails of the underlying probability distribution. These statistical tails are crucial; they represent unique human expressions, idiosyncratic idioms, complex edge cases, and highly specialised domain-specific vocabulary that provide essential depth and genuine creativity to human language. Mathematical models and rigorous simulations conclusively demonstrate that without a continual infusion of authentic, dynamically evolving human-generated content to anchor the system, the overall variance of the training data distribution shrinks. The model's internal representations undergo homogenisation until it eventually collapses into producing highly repetitive nonsense, converging on a narrow band of highly probable but semantically meaningless token combinations. This degenerative process represents a compounding loss of informational fidelity: while initial iterations may retain high accuracy, each subsequent generation suffers from a progressive contraction of the data distribution, which amplifies statistical noise and introduces irreversible mathematical errors.

The process of this algorithmic degradation typically consists of three sequentially identifiable phases, each characterised by an escalating type of structural failure mode. In the first stage, known as the phase of knowledge preservation, models maintain a relatively high degree of operational accuracy and thematic coherence by relying primarily on the robust legacy of verified human data ingested during initial pre-training. However, the transition to the second stage involves the onset of knowledge collapse, a perilous state described in AI literature as the "valley of dangerous competence." During this intermediate phase, models paradoxically remain highly fluent, syntactically perfect, and structurally coherent, which severely masks their underlying epistemological deterioration. They become confidently wrong, consistently producing plausible-sounding but factually untethered outputs, a phenomenon commonly referred to as hallucinations. This phase is particularly hazardous because the remarkably high linguistic quality actively tricks end-users into trusting fundamentally flawed or entirely fabricated information. Finally, in the third and terminal stage of instruction-following collapse, the degeneration becomes absolute and irreversible. Factual reliability plummets to random stochastic baselines, and the underlying mathematical vector representation of complex information breaks down entirely. The model fundamentally loses its basic capacity to comprehend user prompts, resulting in malformed text fragments and an inability to maintain conversational logic [2].

The practical impact of this shift is becoming measurable across major digital platforms. For instance, Wikipedia is currently experiencing a rapid decline in human visitors and active contributors. Similarly,

participation on Stack Overflow — a specialised global forum where programmers collaboratively solve technical problems and share verified code — has regressed to lows not seen since the platform was founded. This abandonment of verified knowledge repositories indicates that users are moving away from human peer-review toward automated AI chatbots. However, when these crucial sources are replaced by machine-generated outputs, the verified data absolutely required to train robust models ceases to exist. This dynamic has led to a severe trust gap, where recent surveys indicate that only 29% of developers highly trust machine-generated code. This erosion of community trust creates a perilous paradox: as corporations increasingly deploy AI to boost short-term productivity, the underlying foundation of verifiable human expertise that these systems depend upon is systematically dismantled. If junior developers and researchers rely entirely on generative models for answers, the critical pipeline of novel, peer-reviewed human problem-solving will permanently stagnate, leaving future edge cases completely unresolved.

To bypass this impending data wall, leading researchers, safety alignment teams, and industry pioneers are shifting their strategic focus away from brute-force data scaling toward a meticulous, computationally efficient, and research-first paradigm. Ilya Sutskever, a highly respected visionary figure in the machine learning field, has persuasively argued that the era of relying solely on massive, uncurated data ingestion is rapidly drawing to a close. He refers to authentic human data as a finite fossil fuel for AI development that is rapidly being depleted [3]. In response, his proposed framework focuses on the advanced concept of weak-to-strong generalisation. In this sophisticated framework, significantly smaller, highly curated, and strictly human-verified models are actively deployed to supervise, align, and correct the outputs of much more powerful next-generation neural networks [4]. This layered methodology allows the overarching AI system to carefully elicit latent knowledge and systematically filter out creeping synthetic errors, ensuring stability even when operating within a heavily polluted, post-collapse internet training environment. Furthermore, another highly promising solution involves moving away entirely from passive text-based data ingestion. Instead, researchers are pioneering interactive, dynamic training methodologies embedded within strictly rule-based virtual environments. These digital training grounds provide rigid, objective feedback based on immutable logic, formalised mathematics, and simulated physics, forcing models to verify facts independently. By embedding AI agents inside these procedurally generated engines, developers can create virtually infinite streams of pristine, non-linguistic training data. This radical shift from static text analysis to active, environment-based reinforcement learning helps mitigate the risks of recursive homogenisation and accelerates general reasoning.

In conclusion, the emergence of model collapse represents a fundamental existential challenge to the prevailing scaling laws of generative artificial intelligence and the entire trajectory of the technology. The ongoing transition toward a synthetic-dominated digital ecosystem necessitates an immediate industry-wide pivot from simplistic quantity-driven scaling to high-fidelity research, rigorous data provenance tracking, and domain-specific human curation. Actively preserving the purity of human-authored datasets and rapidly implementing advanced supervision frameworks serve as essential structural safeguards for the long-term reliability and operational viability of these powerful systems. Furthermore, ensuring the factual integrity of AI outputs is intrinsically tied to the economic viability of the broader tech industry. A failure to mitigate model collapse could trigger widespread loss of enterprise trust, massive financial instability, and severe market recalibrations as the utility of AI tools sharply drops. To avoid this, researchers must prioritise creating isolated data archives strictly free from synthetic outputs. Institutions must also establish international standards for data watermarking to securely differentiate original human intellect from algorithmic approximations. Ultimately, the industry must prioritise the establishment of “human data sanctuaries” to safeguard authentic linguistic complexity, ensuring that future generative models remain anchored by genuine human reasoning rather than drowning in a sea of low-entropy synthetic noise. Without aggressively addressing this recursive feedback loop, the industry risks a systemic failure in automated reasoning capabilities, a tragic devaluation of global human knowledge, and the bursting of an AI-driven economic bubble [5].

References:

1. Shumailov, I. AI models collapse when trained on recursively generated data / I. Shumailov, Z. Shumaylov, Y. Zhao [et al.] // Nature. – 2024. – URL: https://www.nature.com/articles/s41586-024-07566-y (date of access: 12.03.2026).

2. Keisha, F. Knowledge Collapse in LLMs: When Fluency Survives but Facts Fail / F. Keisha, Z. Wu, Z. Wang [et al.]. – arXiv preprint arXiv:2509.04796, 2025. – URL: https://arxiv.org/abs/2509.04796 (date of access: 12.03.2026).

3. Sutskever, I. Data Is the 'Fossil Fuel' of A.I. / I. Sutskever // Observer – 2024. – URL: https://observer.com/2024/12/openai-cofounder-ilya-sutskever/ (date of access: 12.03.2026).

4. Weak-to-strong generalization // OpenAI – 2023. – URL: https://openai.com/index/weak-to-strong-generalization/ (date of access: 12.03.2026).

5. How An AI Bubble Burst Could Shake Global Financial Markets // Oliver Wyman – 2026. – URL: https://www.oliverwyman.com/our-expertise/insights/2026/jan/ (date of access: 12.03.2026).