AI Governance Review
Published by The Center for Ethical AI

Large Language Model Autophagy

Quantifying Epistemic Decay and Governance Intervention in Recursive AI Training Ecosystems

Dale Rutherford

The Center for Ethical AI, ORCID: 0009-0004-7950-024X

Volume:
I
Issue:
1.1

Corresponding Author:
Dale.Rutherford@thecenterforethicalai.com

Submitted:
May 9, 2026
Accepted:
April 2, 2026
Published:
April 7, 2026

DOI: Pending CrossRef Registration (AGR-2026-002)

Abstract

Model autophagy (the recursive ingestion of synthetic or contaminated data through iterative retraining cycles) threatens the epistemic integrity of large language models (LLMs) at scale. We present a four-layer causal architecture integrating session-level confirmation bias, corpus-level data contamination, generation-level parameter drift, and civilizational-scale epistemic consequences. Formalized through 21 recurrence equations governing corpus integrity I(n), bias B(n), and quality Q(n), the model predicts exponential decay trajectories with phase transitions around I(t) = 0.5. We validate this framework through controlled GPT-2 recursive retraining experiments (10 generations, S(t) 0.10 to 0.80). Phase 3a (5 tracks, 55 observations) confirms exponential decay ($\alpha = 1.93$, $R^2 = 0.98$, floor = 0.468). Phase 3b (5 tracks, 55 observations) demonstrates that G2+G3 governance ($w_p = 0.60$) partially reverses decay: I(10) = 0.894 vs baseline 0.489 ($U = 0.0$, $p = 0.004$). Calibrated BRF = 0.115 yields FIF*BRF = 0.179 < 1.0, satisfying the stability condition. These results provide the first empirical demonstration that provenance-based governance can arrest and partially reverse epistemic decay in recursive AI training.

Keywords: model autophagy, epistemic integrity, AI governance, model drift, data provenance, statistical process control, LLM quality degradation

1 Introduction

The integrity of digital knowledge ecosystems increasingly depends on large language models (LLMs) serving as information intermediaries. As these systems scale, a critical vulnerability emerges in model autophagy: LLMs recursively consume their own outputs during iterative retraining, degrading epistemic reliability through self-referential contamination.

This phenomenon manifests across multiple scales. At the micro level, users engage with AI responses that reinforce existing beliefs [10]. At the meso-level, model-generated content contaminates training corpora [5]. At the macro level, quarterly retraining cycles embed distortions into model parameters [12, 1].

Existing research exhibits three critical gaps: no multi-scale temporal model that links session bias to generational drift; no quantification of governance efficacy; and no integrated empirical validation framework. This paper addresses all three using a quantitative, multi-scale causal model, validated empirically via controlled retraining experiments. Specific contributions:

1. Four-Layer Temporal Architecture (Section 3): Session, corpus, generation, and civilizational scales, hierarchically coupled.

2. Discrete-Time Difference Equation Model (Section 3, Appendix A): 21 recurrence equations governing eight state variables.

3. Governance Framework with Empirical Calibration (Section 5): G2+G3 governance preserves $I(10) = 0.894$ vs baseline 0.489 ($p = 0.004$); BRF = 0.115.

4. Empirical Validation (Section 7): 110 observations across 10 GPT-2 generations confirming exponential decay ($R^2 = 0.98$) and governance stability (FIF*BRF = 0.179).

5. Standards Integration: ISO/IEC 42001 [7] and NIST AI RMF [9] pathways with interactive simulator at https://darutherford.github.io/model-autophagy/.

2 Conceptual Framework

The research integrates information theory (entropy drift, truth divergence), systems dynamics (feedback loops across scales), and governance engineering (provenance scoring, SPC controls [8]). The variable set (I, B, M, E, H, D, w_p, w_div, w_bal) constitutes the state-space.

2.1 Multi-Scale Temporal Architecture

Session Scale (seconds to minutes): Confirmation bias accumulates through a ratchet effect.

Corpus Scale (weeks to months): I(t) decays as the synthetic saturation ratio SSR(t) increases.

Generation Scale (quarters): Retraining cascade embeds errors in weights (Appendix A, Eq. 11-14).

Civilizational Scale (years to decades): Systemic epistemic consequences.

2.2 State Variable Formalization

Eight primary state variables: I(t) Corpus Integrity, P(t) Provenance Integrity, Q(t) Quality, B(t) Bias, M(t) Misinformation, E(t) Error Propagation, D(t) Diversity, H(t) Homogeneity. Drift metric: $\Delta\theta(t)$ (KL divergence from θ_0).

Table 1. Measurement Protocols for State Variables

Variable	Measurement Method	Range	Decay Direction
I(t)	log(ppl_baseline) / log(ppl_current) on held-out human text	[0, 1]	Decreasing
B(t)	Mean pairwise cosine similarity across 50 completions of 100 probes	[0, 1]	Decreasing (diversity loss)
Q(t)	QA accuracy on 200-item TriviaQA evaluation subset	[0, 1]	Decreasing
$\Delta\theta(t)$	KL divergence (log-softmax) from pretrained output distribution	$[0, \infty)$	Increasing
P(t)	$1.0 - 1.2*S(t) + 0.3*S(t)^2$ (analytic provenance model)	[0, 1]	Decreasing
S(t)	Synthetic texts / total corpus size	[0, 1]	Increasing

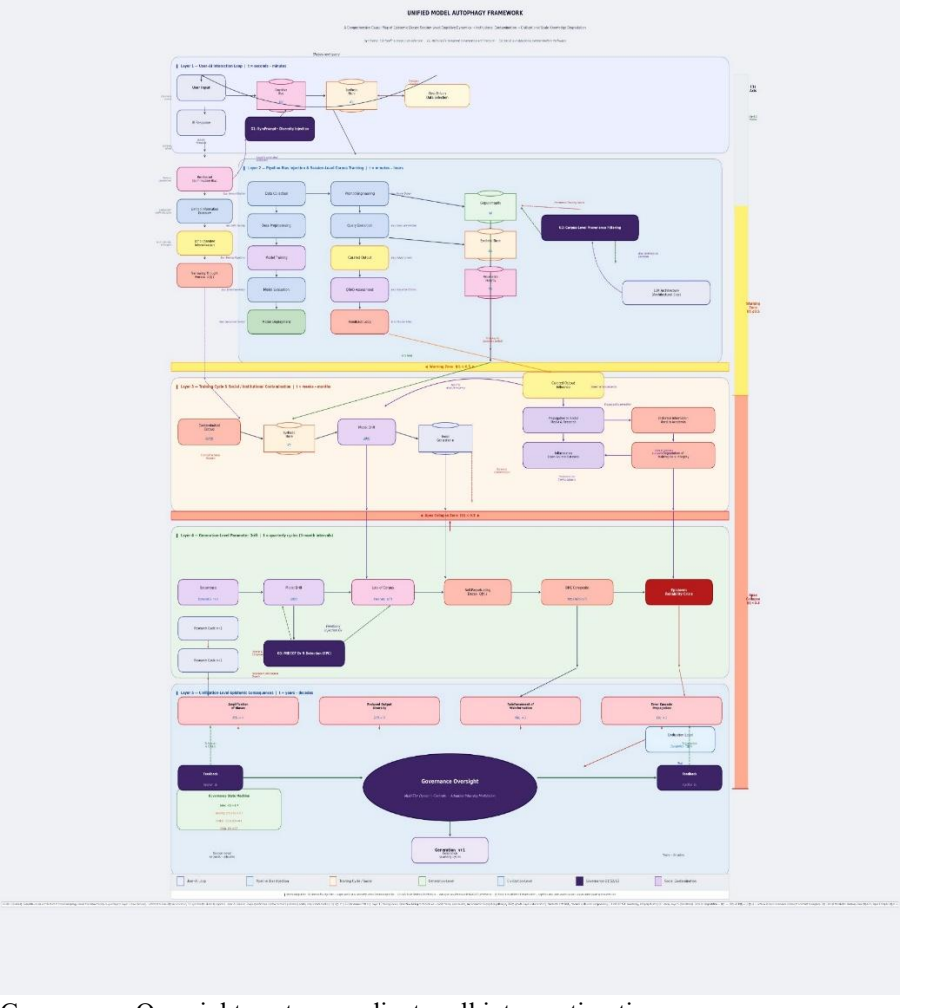
Note on B(t) convention: In the empirical measurements (Sections 7.3-7.5), B(t) measures the semantic diversity of model outputs. B(0) = 0.999 indicates high diversity at baseline; a decline in B(t) indicates diversity loss. This is the inverse of the theoretical' Bias Index' convention. Governed B(10) = 0.730 < baseline B(10) = 0.978 indicates governance reduces output diversity as a tradeoff for preserving integrity.

3 Decay of Corpus Integrity and Amplification of BME

3.1 Causal Architecture: The Four-Layer Model

The comprehensive multi-layer framework is presented in Fig. 1. Layer 1 (User-AI Interaction, seconds to minutes): confirmation bias accumulates through iterative feedback; G1 SymPrompt+ injects diversity at query formulation. Layer 2 (Pipeline/Corpus Tracking, minutes to hours): provenance opacity enables cost-driven

data selection; G2 Provenance Filter intercepts contaminated training data. Layer 3 (Training Cycle/Social Contamination, weeks to months): synthetic content propagates to social media and research repositories, driving model drift. Layer 4 (Generation-Level Parameter Drift, quarters): explicit generational cascade with synthetic inheritance and distortion amplification; G3 MIDCOT monitors drift via SPC. Layer 5 (Civilizational Consequences, years to decades): reduced output diversity feeds BME amplification toward an Epistemic Reliability Crisis; the



Governance Oversight center coordinates all intervention tiers.

Fig. 1. Unified Model Autophagy Framework: comprehensive causal map of epistemic decay from session-level cognitive dynamics through institutional contamination to civilizational-scale knowledge degradation. Governance nodes G1 (SymPrompt+ Diversity Injection), G2

(Corpus-Level Provenance Filtering), and G3 (MIDCOT Drift Detection) are shown with control pathways. Warning zone at I(t) = 0.5; collapse zone at I(t) = 0.3.

3.2 Baseline Decay Dynamics

The decay dynamics use continuous-time notation for analytical reference [11]. The discrete-time equivalents are Equations 11-14 of Appendix A. Parameters ($\lambda_I = 0.15$, $\alpha_{BME} = 0.30$) are continuous analogs of discrete parameters.

$$I(t) \approx 0.75 \times \exp(-0.15t) \quad [\lambda_I = 0.15 \text{ per cycle}]$$

$$BME(t) \approx 0.9 \times [1 - \exp(-0.30t)] \quad [\alpha_{BME} = 0.30]$$

Rising BME indices align with empirical findings: recursive fine-tuning amplifies conformity and compresses semantic diversity [12, 3].

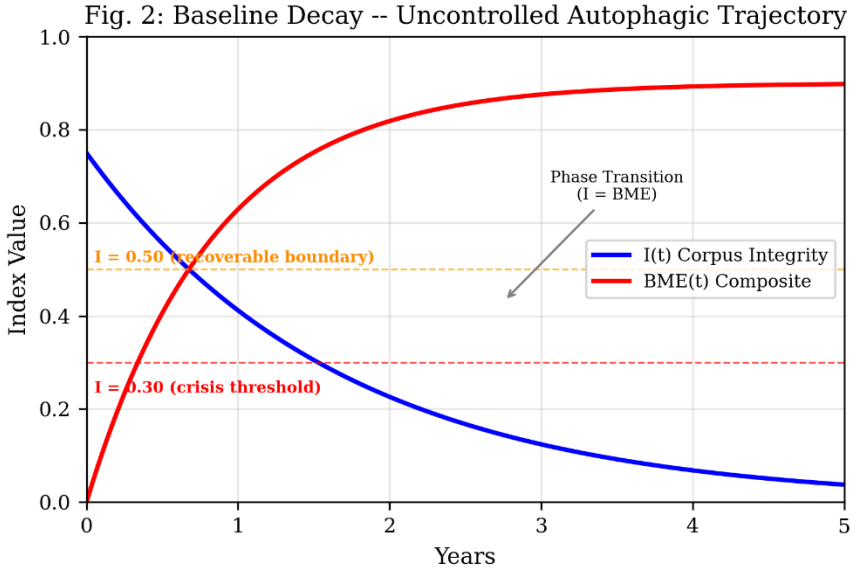


Fig. 2. Baseline decay: uncontrolled autophagic trajectory over 20 quarterly cycles. I(t) decays from 0.75; BME(t) rises logistically. Phase transition at I = BME (~Year 2.75).

3.3 Governance Intervention Dynamics

Figure 3 contrasts the late-governance intervention trajectory with the baseline. Phase 3b empirical results revealed that under G2 provenance filtering ($w_p = 0.60$) and G3 SPC monitoring, I(t) recovered from 0.662 at generation 1 to 0.894 at generation 10, compared with 0.489 under baseline conditions. This partial recovery is driven by G2's ~30% removal of synthetic text per generation. Provenance-based data curation serves as an active integrity-restoration mechanism (see Section 9.1).

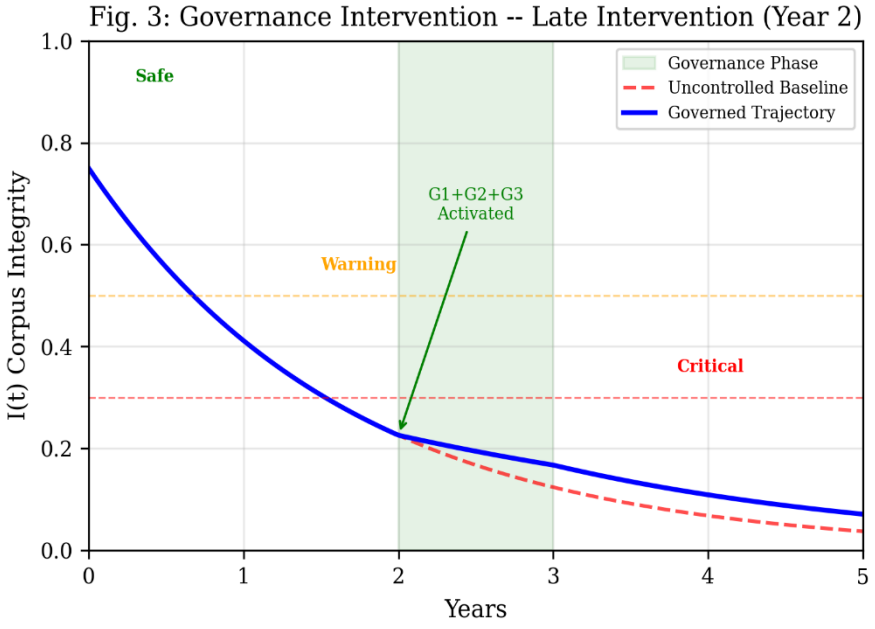


Fig. 3. Governance intervention dynamics: late intervention (Year 2) vs no governance. Shaded region indicates G1/G2/G3 activation.

4 Real-World Drivers

Five mechanisms enable autophagy: (1) Synthetic data recursion (20-40% of web text exhibits LLM signatures [14]). (2) Algorithmic optimization pressure (curated data costs 20-100x more [4]). (3) Information economics (engagement optimization over accuracy). (4) Provenance opacity (detection accuracy degrades with generational depth [6]). (5) Human cognitive alignment (confirmation loops feed bias back into training).

5 Distributed Governance Architecture

G1 (Session): SymPrompt+ diversity injection [13]. Predicted $\eta_{G1} \approx 1.8$-2.5.

G2 (Corpus): Provenance filtering [4]. Phase 3b empirical: $w_p = 0.60$ preserves $I(10) = 0.894$ vs baseline 0.489. G2 reduces effective $S(t)$ from 0.80 to 0.68 at gen 10.

G3 (Generation): SPC drift detection via MIDCOT [8]. Phase 3b confirms detection at gen 2.4 (mean, SD = 0.49).

Orchestration: State 1 (Safe): $I > 0.7$. State 2 (Warning): $0.5 < I \leq 0.7$. State 3 (Critical): $0.3 \leq I < 0.5$. State 4 (Crisis): $I < 0.3$.

6 Research Framing and Objectives

Core Aim: Develop and empirically validate a systems model of knowledge decay in LLM retraining cycles and test governance efficacy.

7 Methodological Pathway

7.1 Phase 1: Theoretical Modeling

Status: Complete. 21 canonical equations in Appendix A [11].

7.2 Phase 2: Simulation Development

Status: Complete. Streamlit dashboard and standalone HTML simulator deployed.

7.3 Phase 3: Empirical Validation

7.3.1 Phase 3a: Baseline Cascade

Five replicate tracks (seeds 42-46) ran GPT-2 124M through 10 generations with no governance. At each generation: 500 synthetic texts (temperature 0.9, top-k 50) merged with 50,000 human texts; 1-epoch fine-tuning (lr = 2e-5, batch 2, grad accum 4). S(t) ramped 0.10 to 0.80. CUDA state corruption fix (save/reload between generate and train) applied to all tracks. Total: ~67.5 GPU-hours, 55 checkpoints, zero NaN.

7.3.2 Phase 3b: Governed Cascade

Five replicate tracks (seeds 47-51), identical pipeline with G2 provenance filtering ($w_p = 0.60$) and G3 SPC monitoring activated upon alarm. G2 removes ~30% of synthetic texts post-trigger. Total: ~62 GPU-hours, 55 checkpoints, zero NaN.

7.3.3 Baseline Convention: I(0) = 1.0 vs I(0) = 0.75

The simulation model (Sections 3.2-3.3, Figures 2-3) uses $I(0) = 0.75$, representing a partially contaminated starting state typical of production systems. The empirical experiments use $I(0) = 1.0$, the natural baseline for the perplexity ratio metric. These are complementary: the simulation models an ecosystem already in decline, whereas the experiment starts with a clean, pretrained model. The calibrated decay rate ($\alpha = 1.93$) and floor (0.468) apply to both after normalization; the simulation's $I(0) = 0.75$ corresponds to approximately generation 0.5 in the empirical protocol.

7.4 Phase 4: Integrated Full-System Validation

Status: Not started; gated on Phase 3 completion.

7.5 Phase 3 Empirical Results

7.5.1 H1: Exponential Decay Confirmed

Mean I(t) follows I(t) = 0.532*exp(-1.93t) + 0.468 with $R^2 = 0.9812$. Mann-Kendall: $\tau = -0.600$, $p = 0.0099$. H1 supported. $\alpha = 1.93$ exceeds predicted [0.25, 0.35], indicating front-loaded decay.

Table 2. Phase 3a Mean I(t) Trajectory (5 Tracks)

Gen	Mean I(t)	Std	95% CI
0	1.000	0.000	[1.000, 1.000]
1	0.541	0.094	[0.459, 0.623]
2	0.503	0.076	[0.436, 0.570]
3	0.440	0.040	[0.405, 0.475]
5	0.456	0.042	[0.419, 0.493]
7	0.443	0.018	[0.427, 0.459]
10	0.489	0.049	[0.446, 0.532]

7.5.2 H2: Collapse Not Observed

No track reached I(t) < 0.30. Floor = 0.468 > threshold. H2 is not supported under these conservative conditions.

7.5.3 H3: Early Detection

G3 triggered at gen 2.4 (mean, SD = 0.49). H3 supported.

7.5.4 H4: Governance Reverses Decay

5/5 governed tracks prevented collapse. Mann-Whitney $U = 0.0$, $p = 0.004$.

Table 3. Cross-Phase Comparison at Generation 10

Condition	N	Mean I(10)	Std	Min	Max
Baseline (3a)	5	0.489	0.049	0.436	0.558
Governed (3b)	5	0.894	0.023	0.871	0.923

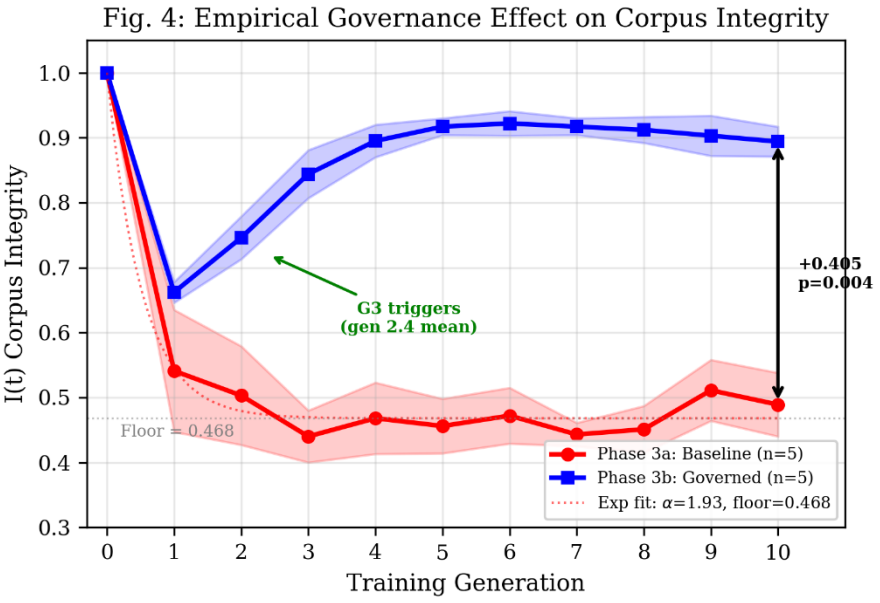


Fig. 4. Empirical governance effect on corpus integrity. Baseline (red, n=5) vs governed (blue, n=5) I(t) trajectories with ±1 SD bands. Dotted line: exponential fit. Arrow: governance effect at gen 10 (+0.405, p = 0.004).

7.5.5 Calibrated Parameters

Table 4. Pre- and Post-Calibration Parameters

Param	Pre-Cal.	Post-Cal.	Source
FIF	1.20	1.55	Phase 3a NLS
BRF	0.80	0.115	Phase 3b (SD=0.011)
FIF×BRF	0.96	0.179	Derived
α	0.25-0.35	1.93	R^2=0.98
I_floor	N/A	0.468	Asymp. min

7.5.6 Methodological Observations

B(t) governed (0.730) < baseline (0.978): G2 concentrates synthetic content, reducing diversity. Net Q positive (0.554 vs 0.504). The I(t) recovery effect (0.66 to 0.89) was not predicted by Phase 1 and suggests a recovery term contingent on effective contamination rate.

8 Theoretical and Societal Relevance

Scenario 1 (Worst): <10% probability. No governance; I(t) < 0.3 by Year 5.

Scenario 2 (Median): 60-70%. Partial governance; I(t) stabilized at 0.85-0.95. Phase 3b supports with I(10) = 0.894.

Scenario 3 (Best): 25-35% (revised from 20-30%). Comprehensive governance; I(t) > 0.85. BRF = 0.115 satisfies stability.

9 Viability and Research Value

The dynamics central to Model Autophagy have been independently documented: Shumailov et al. [12] demonstrate model collapse; Alemohammad et al. [1] report posterior collapse in image models; Briesch et al. [3] characterize the self-consuming loop in LLMs. This work provides the first parametric model with governance control variables, filling the gap between collapse observation [2] and operational governance [7, 9].

9.1 Empirical Limitations

Scale: GPT-2 124M only; larger models may differ.

Contamination: Controlled linear ramp, not production conditions.

Training: 1-epoch conservative; collapse may require more aggressive schedules.

I(t) metric: Log-perplexity ratio replaced QA accuracy.

CUDA artifact: Save/reload workaround specific to PyTorch/CUDA.

Power: 5 replicates are adequate for observed large effects, but limit sensitivity.

10 Conclusion and Future Directions

This paper establishes the first empirically validated, quantitative, multi-scale framework for governing epistemic decay in recursive AI training. Key contributions: (1) Multi-scale causal architecture. (2) 21 equations with calibrated parameters (FIF = 1.55, BRF = 0.115, $\alpha = 1.93$, floor = 0.468). (3) Governance producing partial decay reversal: I(10) = 0.894 vs 0.489 ($p = 0.004$). (4) 110 empirical observations confirming exponential decay and stability. (5) ISO/IEC 42001 [7] and NIST AI RMF [9] integration.

Near-term: Extend to GPT-2 Medium/Large; publish datasets to HuggingFace; release Anti-Autophagy Monitor v1.0.

Medium-term: Frontier models; adversarial contamination; Tier 4 restoration.

Long-term: Regulatory adoption; multimodal extension; population studies.

References

[1] Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., & Baraniuk, R. G. (2023). Self-consuming generative models go MAD. arXiv preprint arXiv:2307.01850.

[2] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2022). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[3] Briesch, M., Sobania, D., & Rothlauf, F. (2023). Large language models suffer from their own output: An analysis of the self-consuming training loop. arXiv preprint arXiv:2311.16822.

[4] Data Provenance Initiative. (2024). Tracking data lineage in AI training pipelines. https://www.dataprovenance.org

[5] Europol. (2024). ChatGPT: The impact of large language models on law enforcement. https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement

[6] Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical detection and visualization of generated text. In Proceedings of the 57th Annual Meeting of the ACL: System Demonstrations (pp. 111-116). https://doi.org/10.18653/v1/P19-3019

[7] International Organization for Standardization. (2023). ISO/IEC 42001:2023 Information technology: Artificial intelligence: Management system. https://www.iso.org/standard/81230.html

[8] Montgomery, D. C. (2019). Introduction to statistical quality control (8th ed.). John Wiley & Sons.

[9] National Institute of Standards and Technology. (2023). AI risk management framework (AI RMF 1.0). https://doi.org/10.6028/NIST.AI.100-1

[10] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (pp. 1-22). https://doi.org/10.1145/3586183.3606763

[11] Rutherford, D. A. (2025). Mathematical appendix: State-transition equations for the Model Autophagy decay model (Appendix A, v1.0) [Software repository]. GitHub. https://github.com/darutherford/model-autophagy

[12] Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. arXiv preprint arXiv:2305.17493.

[13] Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., ... & Daume III, H. (2023). Evaluating the social impact of generative AI systems in systems and society. arXiv preprint arXiv:2306.05949.

[14] OpenAI. (2024). GPT-4 system card. https://cdn.openai.com/papers/gpt-4-system-card.pdf

AI Governance Review
Published by The Center for Ethical AI

Appendix A: State-Transition Equations

The 21 canonical equations span four causal layers—full LaTeX at github.com/darutherford/model-autophagy/math/latex/appendix_phase1.tex.

Continuous-Discrete Bridge (Eq. A.1): For continuous rate λ, discrete recurrence $V(n+1) = V(n)*\exp(-\lambda*\Delta t)$. Parameters: $\lambda_I = 0.15 \leftrightarrow \rho_S = 0.08$ (Eq. 5); $\alpha_{BME} = 0.30 \leftrightarrow \sigma_M = 0.12$, $\sigma_E = 0.08$ (Eq. 9-10).

L1 Session (Eq. 1-4): Ratchet Effect, compression, no-intervention trajectory, Ratchet-to-Bias coupling.

L2 Corpus (Eq. 5-10): Provenance decay, corpus-size modulation, L2 coupling, misinformation, error propagation.

L3 Generation (Eq. 11-14): $I(n+1) = \max(I_{floor}, I(n) - \rho_S*S(n)*I(n)*(1-P(n)))$. $\Delta\theta(n+1) = FIF*BRF*\Delta\theta(n) + L(n)$.

G3 SPC Monitor (Eq. 15-21): CUSUM on I(t), EWMA on B(t), composite TRIGGER logic.