Anti-Autophagy Monitor

User Guide v1.0

Model Autophagy Research Program

Dale A. Rutherford, Ph.D.

1. What Is This Simulator?

The Anti-Autophagy Monitor is an interactive dashboard that demonstrates how large language models (LLMs) degrade when iteratively trained on their own outputs, a phenomenon called model autophagy. The simulator projects the trajectory of six key state variables across multiple retraining generations, comparing an uncontrolled baseline against a governed scenario with user-adjustable intervention parameters.

The simulator is built on a formal 21-equation mathematical model validated through controlled GPT-2 retraining experiments. The corpus integrity decay trajectory (the primary metric) is empirically calibrated with R-squared = 0.98. Governance efficacy parameters are theoretical projections pending validation in Phase 3b.

2. Dashboard Layout

2.1 Sidebar (Left Panel)

The left panel contains all adjustable parameters organized into three sections:

- **Simulation Scope:** Controls the number of retraining generations and the peak synthetic data share. “Generations” sets the number of quarterly retraining cycles to simulate (8 to 40). “Peak Synthetic Share” sets the maximum proportion of model-generated content in the training corpus.
- **Governance Parameters:** Controls when and how strongly governance intervenes. Includes intervention generation timing and three governance weight sliders (w_p, w_div, w_bal).
- **Model Parameters:** Controls the underlying decay dynamics, including FIF (Feedback Injection Factor), BRF (Bias Reduction Factor), and beta_P (provenance amplifier).

2.2 Main Panel (Right)

The main panel displays results in four sections:

- **Disclaimer Banner:** Blue box at the top clarifying which parameters are empirically calibrated and which are projected.
- **Metric Cards:** Six cards showing the governed scenario final-generation values for I(t) Integrity, P(t) Provenance, B(t) Bias, Q(n) Quality, M(t) Misinformation, and E(t) Error. Color-coded green/yellow/red.
- **Delta Row:** Shows the governance effect (governed minus baseline) for each metric. Green indicates governance helped; red indicates it worsened.
- **Chart Tabs:** Four interactive Plotly charts with baseline (red) and governed (blue) scenario traces.

3. Understanding the State Variables

Variable	Description	Range	Good Direction
I(t)	Corpus Integrity. Measures how much of the training data retains original human-authored provenance.	0.0 to 1.0	Higher is better
P(t)	Provenance Integrity. Tracks whether the lineage of training data can be verified back to human sources.	0.0 to 1.0	Higher is better
B(t)	Bias Index. Measures output distribution narrowing (reduced diversity in model responses).	0.0 to 1.0	Lower is better
Q(n)	Quality Index. Composite measure of output quality degradation.	0.0 to 1.0	Higher is better
M(t)	Misinformation Index. Tracks the propagation of factual errors through the training cycle.	0.0 to 1.0	Lower is better
E(t)	Error Propagation. Measures compounding of errors across retraining generations.	0.0 to 1.0	Lower is better

4. The Four Chart Tabs

4.1 Integrity and Quality

This is the most empirically grounded chart. The orange diamond markers show actual Phase 3a experimental data (mean of 5 GPT-2 replicate tracks with 95% confidence intervals). Two reference lines mark the empirical floor at I = 0.468 and the collapse threshold at I = 0.3. The red traces show the uncontrolled baseline; blue traces show the governed scenario.

Key insight: Governance can raise the effective integrity floor above 0.468, but cannot reverse decay that has already occurred.

4.2 Bias and Provenance

Shows B(t) as solid lines and P(t) as dashed lines. B(t) starts near 1.0 (high bias) and declines slowly. P(t) decays as synthetic content dilutes data lineage. Orange diamonds show empirical B(t) data. The governed scenario (blue) maintains higher provenance than the baseline (orange dashed) after the governance activation point.

4.3 Contamination

Displays S(t) synthetic share (amber), M(t) misinformation (solid), E(t) error propagation (dashed), and BME composite (dotted). These metrics are model projections, not yet empirically calibrated. The governed scenario shows how w_div suppresses M(t), and w_bal suppresses E(t).

Note: *The M(t), E(t), and BME traces are theoretical projections with illustrative parameter values. Their specific trajectories will be validated in Phase 3b.*

4.4 Governance State

Classifies system state as Safe ($I > 0.7$), Warning ($0.5 < I < 0.7$), Critical ($0.3 < I < 0.5$), or Crisis ($I < 0.3$). The chart shows state transitions over time for both scenarios. Under default parameters, the baseline typically reaches the Critical state, while governance remains at Warning or Safe.

5. Using the Governance Sliders

All sliders update the simulation automatically with a 200ms debounce. You can also click "Run Simulation" to force a refresh. The charts, metric cards, and delta row all update together.

5.1 Intervention Generation

Controls when governance activates (shown as the purple "Gov. On" marker on all charts). Earlier intervention gives governance more time to act, but it may not be realistic in practice. The slider shows both the generation number and the corresponding year (quarterly cycles, so generation 8 = Year 2.0).

5.2 w_p (Provenance Filter)

Controls the strength of the G2 corpus-level provenance filtering tier. Higher values mean stricter data lineage requirements. Effects:

- Raises the provenance floor (P cannot decay below a w_p-dependent minimum)
- Raises the integrity floor (I stabilizes at a higher value)
- Indirectly reduces M(t) and E(t) by keeping P(t) higher

5.3 w_div (Diversity Weight)

Controls the G1 session-level diversity injection tier. Higher values enforce greater output diversity. Effects:

- Directly suppresses M(t) misinformation propagation
- Reduces E(t) error compounding
- Strongest effect on B(t) bias suppression

5.4 w_bal (Balance Weight)

Controls the human-to-synthetic ratio maintenance. Effects:

- Strongest effect on E(t) error propagation suppression
- Moderate effect on B(t) bias
- Modest raise to the integrity floor

5.5 FIF and BRF

FIF (Feedback Injection Factor): Controls how strongly each generation amplifies drift from the previous one. Default 1.55 is the Phase 3a empirical value. Higher FIF means faster decay.

BRF (Bias Reduction Factor): Controls how effectively governance dampens generation-to-generation drift. The stability condition requires FIF x BRF < 1.0 for convergence. The sidebar indicator shows whether this condition is met.

5.6 Weight Warning

An amber warning appears when w_p + w_div + w_bal > 1.0. This indicates unrealistically strong combined governance. Real-world governance has finite capacity; exceeding 1.0 in aggregate weights represents an idealized scenario that may not be achievable in practice.

6. Empirical Basis and Limitations

Parameter	Status	Source
I(t) decay rate (alpha = 1.93)	Empirical	Phase 3a: 5 GPT-2 tracks, R-sq = 0.98
I(t) floor (0.468)	Empirical	Asymptotic floor from exponential fit
B(t) decline rate (0.002/gen)	Empirical	Mann-Kendall trend, p = 0.0099
FIF (1.55)	Empirical	Phase 3a delta-theta trajectory
M(t) propagation (sigma_M)	Theoretical	Phase 1 model, pending validation
E(t) propagation (sigma_E)	Theoretical	Phase 1 model, pending validation
Governance efficacy (eta)	Theoretical	Pending Phase 3b experiments
Q(t) degradation curve	Theoretical	Logistic on delta-theta, not yet calibrated

6.1 Key Limitations

- Training corpus was locally generated (not real Wikipedia/PubMed data)
- I(t) measured via log-perplexity ratio (revised from QA accuracy during execution)
- Single-epoch fine-tuning may underestimate production decay rates
- Governance slider coupling uses simplified approximations, not the full 21-equation system
- The HTML simulator is a demonstration tool; the Streamlit version (app.py) implements the exact formal equations

7. Suggested Explorations

Try these parameter configurations to understand the dynamics:

Scenario A: No Governance

Set all three governance weights to 0.00 and observe the uncontrolled decay trajectory. I(t) should decline to the floor (0.468) within a few generations.

Scenario B: Maximum Governance

Set w_p = 1.0, w_div = 1.0, w_bal = 1.0. Note the weight warning. Observe how I(t) stabilizes well above 0.6, and M(t)/E(t) are strongly suppressed. This represents an idealized upper bound.

Scenario C: Early vs. Late Intervention

Run with intervention at generation 2 (Year 0.5), then at generation 16 (Year 4.0). Compare the delta rows. Early intervention yields substantially better outcomes because decay is front-loaded (alpha = 1.93 indicates most integrity loss occurs in generations 1 to 3).

Scenario D: Stability Threshold

Adjust BRF downward until the stability indicator turns green (FIF x BRF < 1.0). The threshold is BRF < 0.645. Observe how crossing this threshold changes the governed trajectory from divergent to convergent.

8. CSV Export

Click “Export CSV” to download the complete simulation output for both baseline and governed scenarios. The file contains one row per generation per scenario with all state variables. Column definitions:

Column	Description
scenario	baseline or governed
gen	Generation number (0 to nGen)
year	Calendar year (gen / 4, quarterly cycles)
I	Corpus Integrity Index
P	Provenance Integrity
B	Bias Index
Q	Quality Index
M	Misinformation Index
E	Error Propagation Index
dT	Cumulative parameter drift (delta-theta)
state	System state classification (Safe/Warning/Critical/Crisis)
bme	BME composite (B+M+E)/3

9. Citation and Further Reading

Citation: Rutherford, D. A. (2026). Model Autophagy: Quantifying Epistemic Decay and Governance Intervention in Recursive AI Training Ecosystems. [Manuscript in preparation].

Repository: github.com/darutherford/model-autophagy (private)

Contact: Dale A. Rutherford, Ph.D.

Email: dale.rutherford@thecenterforethicalai.com

Website: https://www.thecenterforethicalai.com/