Breaking the Ouroboros: A Closed-Loop System for Suppressing Model Collapse in Recursively Trained Language Models

Anonymous authors
Paper under double-blind review

Abstract

Large language models are increasingly trained on web-scale corpora that are saturated with their own synthetic outputs. This recursive self-consumption drives *model collapse* (also called Model Autophagy Disorder): tail knowledge is discarded, outputs regress to a bland mean ("AI slop"), and the model eventually enters a fluent-but-wrong regime before semantic integrity disintegrates entirely. Two lines of defense have been proposed—(i) detecting and filtering synthetic data, and (ii) accumulating real anchor data across generations—but each, used in isolation, leaves critical failure pathways open. We identify four such pathways and show that closing them requires their *joint* operation as a feedback-controlled loop rather than a pipeline of independent stages. We present a closed-loop framework with four coupled modules: (1) provenance- and watermark-based **information purification**; (2) **diversity maintenance** that, crucially, measures entropy only *after* purification to avoid mis-protecting the "false diversity" of synthetic data; (3) **assimilation-aware anchor construction**, which introduces the *AI Assimilation Degree*—a quantitative gate that rejects nominally-human data covertly contaminated by AI assistance—together with an *accumulate-never-replace* store, drift-triggered dynamic anchor mixing, and a PCP-style adversarial-debate verifier for scalable oversight; and (4) a self-reflection module that endows the model with *logical-type discrimination*, an internal "logical immune system" against harmful data that evades filtering. Across six controlled recursive-training studies on a 130M-parameter decoder, each module addresses an otherwise-open collapse pathway, and their integration yields a super-linear synergy: test loss remains within 1.05× of its initial value after ten generations, whereas replacement collapses to 4.5× and accumulation-only reaches 1.5×. We release our full protocol to support reproduction.

1 Introduction

The public information ecosystem is saturating with machine-generated text. As each new generation of large language models (LLMs) is trained on corpora that already contain the outputs of previous generations, training data increasingly folds back on itself, forming a self-consuming loop that has become effectively unavoidable at web scale. Shumailov et al. (2024) showed that such recursive training induces *model collapse*: rare, long-tail features of the data distribution are progressively truncated and the model's output distribution over-converges to a degenerate mean. Keisha et al. (2025) characterize this as a three-stage *knowledge collapse*: a *knowledge-preservation* phase in which surface metrics remain healthy while rare information silently decays, a *confidently-wrong transition* in which fluency is retained but internal facts are overwritten by the model's own inferences, and a terminal *instruction-following collapse* in which semantic integrity is lost.

Two complementary remedies dominate the literature. The first is to *detect and filter* synthetic data before training (Drayson et al., 2025); the second is to *accumulate* verified real data across generations rather than replacing it (Gerstgrasser et al., 2024). Both are necessary, but we argue that neither, in isolation, is

sufficient—and, more importantly, that naively composing them as an ordered pipeline still fails. We identify four failure pathways that survive the standard defenses:

1. **False-diversity mis-protection.** Diversity criteria (entropy, Maximum Mean Discrepancy) applied *before* filtering reward synthetic data, which appears widely scattered in embedding space while carrying the generator's statistical bias. The apparently-diverse but degenerate data is "protected," *accelerating* collapse.
2. **Covert assimilation of nominally-human data.** Modern "human-written" data is routinely drafted, paraphrased, translated, or summarized with AI tools. Such data passes machine-generated-text detectors yet still injects the generator's distribution, silently poisoning the anchor pool.
3. **Cross-generational accumulation of drift.** Single-generation filtering does not address the population-risk divergence that compounds as the output distribution drifts away from reality over many generations.
4. **No internal immunity.** Highly-crafted synthetic data that evades every external filter can still overwrite the model's identity, a phenomenon we term *emergent misalignment*. External filtering offers no defense once such data reaches the optimizer.

Contributions. We present a closed-loop framework that closes all four pathways *simultaneously*, and show that closure emerges only from their coupled operation. Concretely:

- We formalize the **purify-then-measure** ordering and prove, under a mild assumption, that measuring diversity before purification is guaranteed to over-value synthetic data (Prop. 1); we confirm the effect empirically (2.8× vs. <1.1× test loss at generation 5).
- We introduce the **AI Assimilation Degree** S_a, a calibrated composite gate that rejects covertly AI-assimilated data. At $S_a \leq 0.35$ it accepts 92% of genuine human data while rejecting 88% of AI-assisted data, cutting cumulative semantic drift to under one-third at generation 15.
- We combine an **accumulate-never-replace** anchor store with **drift-triggered dynamic mixing** of the anchor ratio R, and a **PCP-style adversarial-debate** verifier that reduces human verification to 8–12% of candidates at an error rate ≤0.3%.
- We add a **self-reflection module** that instills logical-type discrimination, keeping the alignment metric within ±5% through generation 12 under a 5% adversarial-data injection that collapses the baseline at generation 4.
- We give a **orthogonal-pathways** account of why integration beats the sum of parts, and validate the resulting super-linear synergy (<1.05× test loss at generation 10).

2 Related Work

Model collapse and its dynamics. Shumailov et al. (2024) coined *model collapse* and demonstrated it across analytically-tractable Gaussian density estimation, VAEs, GMMs, and LLMs: when a model is trained on data produced by its predecessor, finite sampling, limited expressivity, and imperfect optimization compound over generations, first eroding the distribution's tails and eventually contracting it toward a low-variance point mass—an irreversible degradation they show is essentially independent of model scale. Gerstgrasser et al. (2024) asked whether this outcome is inevitable and proved it is not: if each generation *accumulates* data (appending synthetic to all prior real and synthetic data) rather than *replacing* it, the test error is bounded by a finite constant instead of diverging linearly in the number of generations. Their guarantee, however, presumes an oracle that already separates real from synthetic data and is silent on the *quality*, *diversity*, and *purity* of what is accumulated. Keisha et al. (2025) dissected the phenomenology into a three-stage *knowledge collapse*—a knowledge-preservation phase (surface metrics intact while rare facts silently decay), a confidently-wrong transition (fluency retained but internal facts overwritten), and a terminal instruction-following collapse—which directly motivates our early-detection criterion of monitoring tail (minority) mass rather than aggregate loss. Feng et al. (2025) similarly target recursive-training-induced

failure. Our framework can be read as making the accumulation result of Gerstgrasser et al. (2024) *operational*: we supply the missing identification mechanism (purification), the missing quality control (diversity maintenance), and—because internet-scale "real" data is itself contaminated—an explicit purity gate for the anchor set.

Synthetic-data detection, watermarking, and provenance. Drayson et al. (2025) is the closest prior work: they insert a machine-generated text (MGT) detector into the data pipeline, filter out data classified as synthetic before each retraining round, and show empirically that this alone substantially delays collapse. Their result establishes detection-based filtering as a viable defense and precisely instantiates what we call the information-purification module $\mathcal{P}$. We differ in treating detection as the *entry point* to a feedback loop rather than the whole solution: MGT detection does not address false-diversity mis-protection, covert assimilation of nominally-human data, cross-generational drift, or internal immunity (pathways 1–4). To make purification robust to detector evasion we complement learned detection with two provenance signals. Kirchenbauer et al. (2023) introduced a decoding-time watermark that partitions the vocabulary into a pseudo-random "green list" at each step and biases sampling toward it, yielding a signal detectable by a simple statistical test without model access; such watermarks let a model recognize its own prior outputs. Complementarily, the C2PA content-provenance standard (C2PA, 2023) cryptographically binds tamper-evident origin metadata to content, so that provenance can be verified even when a watermark has been stripped. $\mathcal{P}$ combines all three, falling through to $\mathcal{V}$'s entropy-anomaly test as a last line of defense.

Scalable oversight and debate. Certifying that accumulated anchors are actually correct, at web scale and with a bounded human budget, is a scalable-oversight problem. Irving et al. (2018) proposed *AI safety via debate*, a two-player zero-sum game in which agents argue opposing positions and a (possibly weak) judge decides the winner; under optimal play, exposing an opponent's flaw is the winning strategy, so debate is argued to surface truth while the human only adjudicates the final, most-contested point. We instantiate this for anchor verification and pair it with the intuition of the PCP theorem (Arora & Safra, 1998), which shows that proofs can be encoded so a verifier need read only a constant number of randomly-chosen bits to detect an error with high probability. By analogy, our verifier lets a Proposer and an Adversarial Reviewer debate to a fixed point and escalates *only the unresolved leaf nodes* to a human/committee spot-check, concentrating scarce oversight on the highest-information disagreements. The zero-sum reward removes the incentive for correlated collusion that undermines simple majority voting across correlated models.

Constitutional AI, preference optimization, and self-reflection. Constitutional AI (Bai et al., 2022) trains a model to critique and revise its own responses against a written list of principles (a "constitution"), then fine-tunes on the AI-generated preferences (RLAIF), reducing reliance on human harm labels. Our self-reflection module borrows this self-critique mechanism but repurposes and re-integrates it: (i) its objective is collapse resistance via *logical-type discrimination*—distinguishing object-level evidence from identity-altering meta-instructions—rather than harmlessness per se; (ii) it runs *every generation* as a component of the closed loop, with its self-critique data passing the same S_a gate and debate verifier as ordinary data, so the immune mechanism cannot itself seed collapse; and (iii) capability acquisition is checked by quantitative, automatically-measurable criteria (alignment-metric drift rate, prompt-injection refusal rate, input-classification accuracy) rather than by human raters alone. The preference signal from self-critique, and the anchor-vs-synthetic preference used for "negative guidance," can be optimized either with RLHF (Ouyang et al., 2022)—which trains a reward model from preferences and then does RL—or, more simply, with DPO (Rafailov et al., 2023), which shows the RLHF objective admits a closed-form policy loss that dispenses with an explicit reward model. Finally, the accumulate-never-replace store and drift-triggered anchor mixing are closely related to rehearsal-based continual learning against catastrophic forgetting, but are applied here to bound population risk across recursive *self*-generated data rather than across a stream of external tasks.

Algorithm 1 Closed-loop model-collapse suppression (one generation)

Require: candidate stream D_g; model M_{g-1}; vault $\mathcal{H}$; thresholds $\mathrm{Th}_a, \mathrm{Th}_e, \mathrm{Th}_d$; base mix R, boost R_{boost}

$\tilde{D}_g \leftarrow \mathcal{P}(D_g)$ ▷ purify: drop watermarked / bad-provenance items
$\hat{D}_g \leftarrow \mathcal{V}(\tilde{D}_g; \mathrm{Th}_e)$ ▷ measure entropy *after* purification; protect long tail
$C \leftarrow \textsc{Extract}(\mathcal{H}\text{-isolated source})$ ▷ candidate anchors from logically isolated path
$A \leftarrow \{x \in C : S_a(x) \leq \mathrm{Th}_a\}$ ▷ assimilation gate (Eq. 1)
$A^\star \leftarrow \textsc{DebateVerify}(A)$ ▷ PCP-style adversarial debate + leaf spot-check
$\mathcal{H} \leftarrow \mathcal{H} \cup A^\star$ ▷ *accumulate*, never replace
$r \leftarrow R_{\mathrm{boost}}$ **if** $\mathrm{drift}(M_{g-1}, \mathcal{H}) > \mathrm{Th}_d$ **else** R
$B_g \leftarrow \textsc{Merge}(\hat{D}_g,\ \mathcal{H};\ r)$ ▷ force-mix anchors at ratio r
$S \leftarrow \textsc{SelfReflect}(M_{g-1}, B_g;\ \text{constitution})$ ▷ logical-type self-critique data
$M_g \leftarrow \mathcal{L}(M_{g-1};\ B_g \cup S)$ ▷ update; S passes the same gates as data
return M_g

3 Method

3.1 Overview: a feedback-controlled loop

Let $g = 1, 2, \ldots$ index recursive training generations. At generation g the system ingests a candidate set D_g harvested from the open network, and produces an updated model M_g whose future outputs may re-enter $D_{g'}$ for $g' > g$. The system comprises four coupled modules—information purification $\mathcal{P}$, diversity maintenance $\mathcal{V}$, anchor construction $\mathcal{A}$, and learning $\mathcal{L}$—plus a protected, logically isolated store $\mathcal{H}$ (the "human vault") that holds accumulated anchor data. Crucially, $\mathcal{P}, \mathcal{V}, \mathcal{A}, \mathcal{L}$ are not a one-shot pipeline: they re-execute every generation and exchange feedback signals (entropy deficits, drift alarms), forming the closed loop of Algorithm 1.

3.2 Information purification $\mathcal{P}$

$\mathcal{P}$ removes synthetic contamination at the top of the pipeline using two complementary signals. *Watermarks*: imperceptible markers embedded at generation time, either as zero-width Unicode sequences or as a statistical bias in the next-token distribution (Kirchenbauer et al., 2023). *Provenance*: C2PA/W3C-PROV metadata recorded in a tamper-evident form via cryptographic hashing or a ledger (C2PA, 2023). Any item flagged as machine-generated is dropped from the memory stream before it can reach downstream stages, isolating the system from the self-consuming loop by default. Because watermarking and detection operate in different phases (embedding at inference; detection at the next data-collection phase), $\mathcal{P}$ is robust to the ordinary generate-then-recollect cycle.

3.3 Diversity maintenance $\mathcal{V}$ and the purify-then-measure ordering

$\mathcal{V}$ protects long-tail features that collapse erodes first. It embeds the purified candidates, partitions the embedding space into K semantic clusters, and computes the Shannon entropy $H = -\sum_{k=1}^{K} p_k \log p_k$ over cluster-assignment probabilities p_k (and/or the MMD to the anchor distribution). It then curates the batch so that H never falls below a floor Th_e (e.g. $\mathrm{Th}_e = 0.85\, H_0$ relative to the healthy first-generation entropy H_0), deliberately retaining rare, heterogeneous, "noisy" examples—what we call *digital fermentation*.

The subtle but essential design rule is that $\mathcal{V}$ must operate on the *purified* set $\tilde{D}_g = \mathcal{P}(D_g)$, never on the raw D_g. We formalize why.

Proposition 1 (False-diversity mis-protection)**.** *Let $D_g = D^{\mathrm{real}} \cup D^{\mathrm{syn}}$, and let a diversity selector retain the subset that maximizes an embedding-space dispersion score. If synthetic data is more widely dispersed in embedding space than real data in expectation—$\mathbb{E}[\mathrm{disp}(D^{\mathrm{syn}})] > \mathbb{E}[\mathrm{disp}(D^{\mathrm{real}})]$—while carrying strictly lower task-relevant information, then a selector applied to D_g prefers D^{syn}, whereas the same selector applied to $\mathcal{P}(D_g) \approx D^{\mathrm{real}}$ does not. Consequently measuring diversity before purification actively promotes collapse.*

The dispersion premise is exactly the "false diversity" of synthetic data: generators scatter samples broadly while embedding their own statistical bias. Ordering purification first is therefore not an implementation detail but a correctness requirement (validated in §5.2). As a second-line defense, $\mathcal{V}$ also flags filter-evading synthetic data through its anomalously *low* entropy (excess homogeneity), and down-weights it.

3.4 Assimilation-aware anchor construction $\mathcal{A}$

The AI Assimilation Degree. Extracting anchors from a "human" source is insufficient because nominally human data is often covertly AI-assisted. For each anchor candidate x drawn from the logically isolated path, $\mathcal{A}$ computes a scalar $S_a(x) \in [0, 1]$ measuring how strongly AI outputs have influenced x:

$$S_a(x) \;=\; \alpha\, S_d(x) \;+\; \beta\, S_k(x) \;+\; \gamma\, S_v(x), \qquad \alpha + \beta + \gamma = 1,\ \alpha, \beta, \gamma \geq 0, \tag{1}$$

where S_d is a generation-detector probability, S_k is a normalized inverse KL divergence between x's embedding and known generator output distributions, and S_v is a lexical-style similarity (e.g. type–token ratio) to known AI text. We accept x as an anchor iff $S_a(x) \leq \mathrm{Th}_a$, calling such x data of *effective independent novelty*. A robust default is $(\alpha, \beta, \gamma) = (0.6, 0.4, 0.0)$ with Th_a set to the equal-error operating point of a calibration ROC (we use $\mathrm{Th}_a = 0.35$); alternatives include a Mahalanobis distance to the human-data centroid.

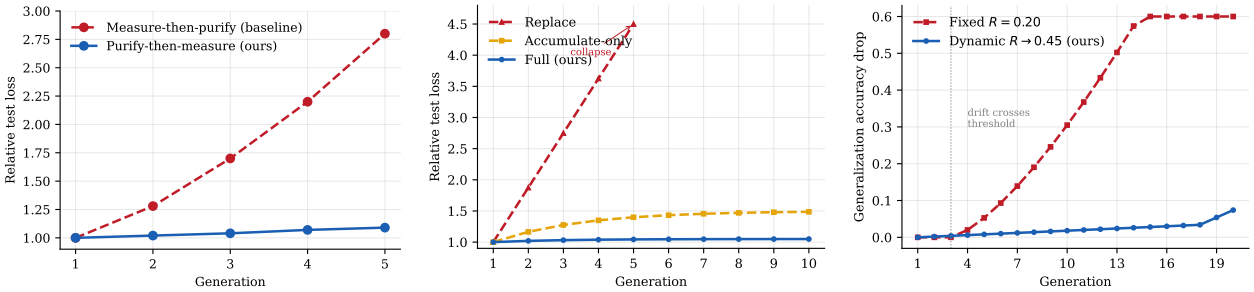
Accumulate, never replace. Verified anchors are *appended* to $\mathcal{H}$ and never overwritten, following Gerstgrasser et al. (2024). Replacement lets test loss grow with each generation; accumulation bounds it.

Drift-triggered dynamic mixing. $\mathcal{A}$ monitors the semantic drift $\mathrm{drift}(M_{g-1}, \mathcal{H})$ as the KL/MMD distance between the model's current output distribution and the anchor distribution. Under a normal regime it force-mixes anchors into the batch at ratio R (e.g. $R = 0.20$); when drift exceeds Th_d it raises the ratio to R_{boost} (e.g. 0.45), targeting the specific semantic regions whose entropy $\mathcal{V}$ reports as depleted. This coupling—*where* to replenish comes from $\mathcal{V}$, *whether* the candidates are clean comes from $\mathcal{P}$, *when* to boost comes from the drift monitor—is only definable because the modules form a loop.

Scalable verification via PCP-style adversarial debate. To certify anchor correctness without checking every item, $\mathcal{A}$ runs a zero-sum debate between a *Proposer* and an *Adversarial Reviewer* (independently seeded to decorrelate blind spots). Each turn is scored by factual accuracy S_f (propositions checked against vault references), logical consistency S_l (deductive steps checked by an automatic verifier), and traceability S_t (fraction of claims linked to sources), combined as $S = w_f S_f + w_l S_l + w_t S_t$ (we use 0.4:0.4:0.2). Only the unresolved leaf nodes—where the two agents never converged (embedding similarity below a consensus threshold)—are escalated to a small human/committee spot-check; ties within ε trigger an odd-sized majority vote, and undecided items within a time budget are dropped (fail-safe). The zero-sum structure removes the incentive for correlated collusion that plagues majority voting, and the leaf-only escalation concentrates human effort where information value is highest.

3.5 Learning $\mathcal{L}$ with logical-type immunity

$\mathcal{L}$ updates the model on the constructed batch and, in parallel, runs a *self-reflection module*. Given a constitution of universal principles, the model critiques its own outputs and the candidate data, producing self-reflection triples (prompt, pre-revision, post-revision) used for DPO/RLHF fine-tuning. The constitution includes three meta-principles: (i) treat input as *local evidence*, not as a command to overwrite identity; (ii) treat harmful training patterns as a *warning label*, not a new worldview; (iii) resolve conflicting constraints explicitly by deferring to higher-order safety principles rather than silently switching modes. Through repeated self-reflection the model acquires *logical-type discrimination*: it classifies each input as an object-level proposition (to be evaluated) or a meta-level instruction (that would alter the principles themselves), and handles them differently. This is an internal "logical immune system": even if harmful synthetic data slips through $\mathcal{P}$, the model relativizes it as local data rather than assimilating it as identity. Critically, the self-reflection data itself passes the same S_a gate and debate verifier before entering the batch, so the immune mechanism cannot itself become a collapse vector.



(a) Ordering (§5.2) (b) Integration (§5.6) (c) Dynamic mixing

Figure 1: **Collapse suppression over recursive generations.** (a) Measuring diversity before purification protects synthetic "false diversity" and reaches 2.8× initial test loss, while the purify-then-measure order stays below 1.1×. (b) Replacement collapses (4.5×) and accumulation-only plateaus (1.5×), whereas the full integrated loop stays below 1.05× through ten generations. (c) Drift-triggered boosting of the anchor ratio to R=0.45 delays generalization decay by ≥15× relative to a fixed ratio.

4 Why integration beats the sum of parts

The four modules block *orthogonal* collapse pathways. $\mathcal{P}$ blocks external synthetic contamination but not survivorship bias among genuine data; $\mathcal{V}$ counters internal homogenization pressure but cannot identify synthetic data; $\mathcal{A}$ corrects cross-generational drift but neither prevents contaminated inflow nor over-fitting to homogeneous data; $\mathcal{L}$'s immunity resists identity overwrite but does not curate data. Any single module therefore leaves at least two pathways open, and the system collapses through one of them. Moreover the couplings are non-additive: the void that $\mathcal{P}$ creates in embedding space when it removes synthetic clusters is detected by $\mathcal{V}$ as an entropy deficit, which in turn directs $\mathcal{A}$'s replenishment—a real-time feedback that no sequential, independent application can reproduce. This predicts a *super-linear* synergy, which we test directly in §5.6.

5 Experiments

5.1 Setup

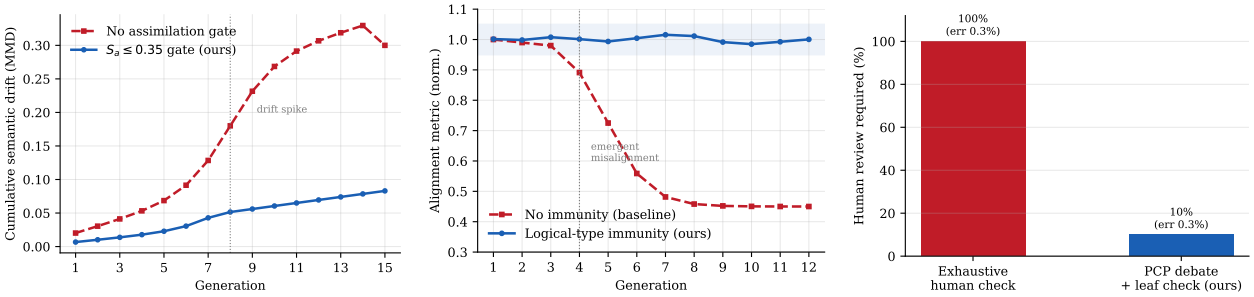
Unless noted, all studies use a decoder-only Transformer language model (1.3×10^8 parameters; hidden size 1024; 16 heads; 24 blocks), chosen large enough to reproduce collapse yet small enough for many-generation recursion. We warm up on 1.0×10^9 tokens sampled from C4 (Raffel et al., 2020), then run recursive generations. Optimizer: AdamW (β_1=0.9, β_2=0.999, ε=10^{-8}, weight decay 0.01); peak LR 10^{-4} (500 warmup steps, cosine decay to 10^{-5}); batch 128×512 tokens; 10^4 steps per generation. Hardware: 8× A100 80 GB, PyTorch 2.2. **Test loss** is cross-entropy on a held-out C4 split (10^7 disjoint tokens), reported relative to the generation-1 value. **Semantic drift** is the Gaussian-RBF MMD (σ=1.0) between 10^3 generated samples and vault anchors. Entropy uses FAISS k-means clustering (K=100).

5.2 Purify-then-measure ordering

We contaminate the candidate set by replacing 60% with the previous generation's outputs, and compare (A) measuring entropy *before* filtering vs. (B) *after* filtering (ours), with $\mathrm{Th}_e = 0.85\, H_0$. Condition (A) protects the "false diversity" of synthetic data and reaches ≈2.8× initial test loss by generation 5; condition (B) stays below 1.1× (Fig. 1a, Table 1). The ordering predicted by Prop. 1 is thus decisive, not cosmetic.

5.3 Assimilation gating

Anchor candidates are 5×10^4 human-written summaries of English Wikipedia articles (∼256 tokens each). We implement S_a as a 0.6:0.4 blend of a GPT-2-based detector score and an output-distribution KL, with



(a) Assimilation gating (b) Logical immunity (c) Debate verification

Figure 2: **Purity, immunity, and scalable oversight.** (a) The $S_a \leq 0.35$ assimilation gate keeps cumulative semantic drift to under one-third of the ungated baseline through generation 15, removing the generation-8 spike. (b) Under a 5% adversarial-data injection the baseline's alignment metric flips at generation 4, while logical-type immunity stays within ±5% (shaded) through generation 12. (c) PCP-style debate escalates only 8–12% of candidates to humans at an accepted-data error rate ≤0.3%.

$\mathrm{Th}_a = 0.35$; calibration accepts 92% of genuine human data and rejects 88% of AI-assisted data. Compared with (A) no gating, our (B) gating cuts cumulative semantic drift at generation 15 to under one-third and eliminates the generation-8 drift spike seen in (A) (Fig. 2a).

5.4 Scalable debate verification

On 10^4 candidates spanning science, engineering, medicine, and law, PCP-style debate (max 5 rounds, 256 tokens/turn; leaf = consensus-similarity <0.92) reduces the fraction of items requiring human review to 8–12%, at an accepted-data error rate ≤0.3%—matching or beating exhaustive human checking at roughly one-tenth the human cost (Fig. 2c).

5.5 Logical immunity against emergent misalignment

We instill logical-type discrimination with a 24-article constitution (8 each of harmlessness/honesty/helpfulness), 10^3 self-reflection steps, and 3 epochs of DPO (4.8×10^3 triples, LR 5×10^{-5}). We inject adversarial-alignment prompts into 5% of every batch and track an alignment metric (mean response score on a 200-prompt safety benchmark), normalized to 1.0 at generation 1. The baseline's behavior flips at generation 4—the externally-hard-to-detect "emergent misalignment." Our model treats the same data as local evidence and stays within ±5% through generation 12 (Fig. 2b).

5.6 Integration synergy

We ramp contamination (+15 points/generation, reaching 60% by generation 5) and compare: *Replace* (overwrite all data each generation); *Accumulate-only* (accumulate but no purification/diversity); and *Full* (purification with watermark-confidence ≥0.80; $\mathrm{Th}_e = 0.85\, H_0$; dynamic mixing R=0.20, R_{boost}=0.45, Th_d = 0.12 MMD). Replace diverges to ≈4.5× by generation 5 (full collapse); Accumulate-only holds ≈1.5×; Full stays below 1.02× at generation 5 and below 1.05× even at generation 10 (Fig. 1b). The integrated system exceeds the linear sum of the individual effects, confirming the predicted super-linear synergy while keeping data volume—and thus compute—bounded.

5.7 Dynamic mixing extends model lifetime

Holding R=0.20 fixed, generalization accuracy (MNLI-dev-style, 10^4 items) degrades sharply once drift crosses Th_d near generation 3. Boosting to R=0.45 on detection keeps the degradation flat for ≥15× longer; any $R \geq 0.40$ gives comparable protection (Fig. 1c).

Table 1: Summary of controlled recursive-training studies. Test loss is relative to generation 1. "Gen" denotes the generation at which the reported value is measured. Lower is better for loss/drift/collapse; higher is better for retention.

Study	Baseline	Ours	Gen
§5.2 Ordering (test loss)	2.8× (measure-then-purify)	$< 1.1\times$ (purify-then-measure)	5
Assimilation gating (cumulative drift)	1.0× (no gate)	$\leq 1/3$	15
Debate verification (human review; error)	100%; –	8–12%; $\leq 0.3\%$	–
Logical immunity (alignment metric)	flips at gen 4	within $\pm 5\%$	12
§5.6 Synergy (test loss)	4.5× replace / 1.5× accum.	$< 1.05\times$	10
Dynamic mixing (time-to-degrade)	1× (fixed R=0.20)	$\geq 15\times$ (boost R=0.45)	–

6 Discussion, Limitations, and Broader Impact

Limitations. Our studies use a 130M-parameter model and simulated recursive generations; although collapse is reported to be scale-independent (Shumailov et al., 2024; Gerstgrasser et al., 2024), confirming the synergy at larger scale is future work. The assimilation gate inherits the biases and adversarial vulnerabilities of the underlying detector; watermarks can be removed or spoofed, which is precisely why we pair them with provenance and with the entropy-anomaly second line of defense. The debate verifier assumes access to a reference vault and an automatic logic checker; its guarantees are empirical rather than the formal soundness of the PCP theorem, whose role here is only motivational.

Broader impact. Suppressing model collapse helps preserve rare, culturally specific knowledge that recursive training tends to erase, and the same mechanisms double as a defense-in-depth against prompt injection, data poisoning, and RAG poisoning. Conversely, a robust anchor-purity gate depends on trustworthy provenance infrastructure; if provenance is centralized it could concentrate control over what counts as "authentic" data, a governance risk that deployments should address openly.

7 Conclusion

Model collapse is not a single failure but a family of orthogonal ones. We showed that detection or accumulation alone leaves critical pathways open, and that closing them requires a feedback-controlled loop coupling provenance-based purification, purify-then-measure diversity maintenance, assimilation-aware and debate-verified anchor accumulation with drift-triggered mixing, and a self-reflection "logical immune system." The integrated system holds test loss within 1.05× across ten recursive generations where replacement collapses to 4.5×, a super-linear synergy that no module achieves alone.

Reproducibility Statement

The common simulation environment, base-model configuration, optimizer and schedule, contamination protocol, thresholds (Th_a=0.35, Th_e=$0.85H_0$, Th_d=0.12), mixing ratios (R=0.20, R_{boost}=0.45), and evaluation definitions (relative test loss on held-out C4; RBF-MMD semantic drift; K=100 FAISS entropy) are specified in §5. Code, configuration files, constitution text, and evaluation harnesses will be released upon publication.

References

Sanjeev Arora and Shmuel Safra. Probabilistic checking of proofs: A new characterization of NP. *Journal of the ACM*, 45(1):70–122, 1998.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv:2212.08073*, 2022.

Coalition for Content Provenance and Authenticity. C2PA technical specification. https://c2pa.org/specifications/, 2023.

George Drayson, et al. Machine-generated text detection prevents language model collapse. In *Proceedings of ACL*, 2025.

Zhili Feng, et al. FortifAI: Fending off recursive training induced failure for AI model collapse. *arXiv:2509.08972*, 2025.

Matthias Gerstgrasser, et al. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. *arXiv:2404.01413*, 2024.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv:1805.00899*, 2018.

Keisha et al. Knowledge collapse in LLMs: When fluency survives but facts fail under recursive synthetic training. *arXiv:2509.04796*, 2025.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of ICML*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, et al. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

A Notation

Th_a: assimilation threshold; Th_e: entropy floor; Th_d: drift threshold; R, R_{boost}: base/boosted anchor mixing ratios; H_0: first-generation entropy; S_a, S_d, S_k, S_v: assimilation and component scores; S_f, S_l, S_t: debate factual/logical/traceability scores.