SEPARATING TRAINING-DATA SOURCE FROM UPDATE RATE IN VERIFIED SELF-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Iterative self-training filters a model's own generated solutions by answer verification and fine-tunes on the survivors. Correct final answers, however, do not guarantee that repeated updates preserve performance, and lowering the learning rate to reduce forgetting also changes the data the model generates in later rounds, so retention differences cannot be attributed to the update or to the data alone. We separate these two factors for Llama-3.1-8B-Instruct over forty rounds on a fixed GSM8K stream. Generator and learner controls first establish that retention depends on the recursive update policy. A crossed design then freezes the generated training data from high-rate and low-rate runs and trains fresh receivers at each rate, with three seeds per cell. At the fixed low receiver rate, data generated under the low-rate policy improves the final-five-round near-domain accuracy change by 7.33 percentage points on average, positive in every seed; the same source contrast is 6.78 points at the fixed high receiver rate and 5.84 points in a separate confirmation on a new split, new seeds and a new server. The same frozen banks in front of a second pre-training family, Qwen2.5-7B-Instruct, reproduce the contrast at two seeds, 4.83 points strict and 4.97 lenient. The fixed-bank receiver-rate contrasts are 0.44 and 1.00 points. A historical 2.33-point round-zero spread comes from a different evaluation setup and does not quantify uncertainty for this crossing. Across six banks, token volume and n-gram diversity do not consistently order with retention, while lexical proximity is higher in the more damaging banks. These descriptions neither identify nor exclude causal mechanisms. Yet every primary run still loses strictly held-out accuracy, and a reference trained on authored rationales also declines in every seed. On GSM8K, verification leaves a data-source component of forgetting that rate tuning does not remove, without useful self-improvement. A completed ARC-Challenge crossing does not reproduce the source contrast at its prespecified measurement resolution.

1 INTRODUCTION

Language models can generate training targets, select which to trust, and update their own weights (Yuan et al., 2024; Zweiger et al., 2025; Zuo et al., 2025). Answer verification looks like a natural safeguard: train only on correct solutions. Yet a correct final answer need not make its rationale a useful target for repeated fine-tuning. We ask how retention depends jointly on generator policy, learner history, and update strength when every retained answer is correct.

We study this question through continual-learning measurements (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2018). An independent-restart control resets both generator and learner at each round. A frozen-generator control retains the accumulating learner but draws every new training set from the unchanged base model. A recursive-reset control retains the current generator but resets the learner each round, completing the factorial test of recursion and accumulation; we call each of the four generator/learner combinations in that factorial a *cell*, and label this one cell R.

The original two-sample, two-epoch experiment finds oracle near degradation of -5.67 ± 1.01 points, despite perfect retained-answer precision. Frozen-generation and learner-reset controls have smaller losses. Crucially, a new matched-runtime comparison shows sensitivity to learning rate: high-rate near changes are $-7.20/-8.67/-8.13$ points, whereas low-rate changes are $-0.73/+1.00/-0.73$. Correct-answer training can therefore have different retention outcomes un-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

der the same verifier and round budget. The contrast is a recursive policy effect: altered updates may change all later generated targets and realized exposure.

The rate comparison alone cannot tell whether its retention difference follows the learner's own update rate or the complete training sequence produced upstream. We therefore freeze the selected training data of each earlier run — we call such a frozen corpus a *bank*, the run that produced it the *donor*, and a freshly initialised learner trained on it a *receiver* — and cross the source condition of the bank with both receiver rates (Figure 1). In the completed primary matrix of this crossed experiment, which we label C, the low-rate receiver has a mean near change of -7.844 points with the high-rate bank and -0.511 with the low-rate bank. The resulting 7.333-point source contrast, positive in every seed, is the C-phase primary result, measured on the original split with seeds $0/1/2$. The same contrast at the fixed high receiver rate is 6.778 points, whereas the two fixed-bank receiver-rate contrasts are 0.444 and 1.000 points, numerically below the historical 2.333-point round-zero spread reported in Section 3.4. That spread comes from a different evaluation setup and is only a descriptive reference, not an uncertainty bound for these contrasts. Learning gains on the held-out training slices are absent, leaving C exploratory.

A completed confirmation separates that contrast from the split, seeds and machine where it was first measured: fifteen further forty-round runs, on a new GSM8K partition with seeds $101/102/103$ and a different server, reproduce it at mean $+5.84$ points, positive in every seed. That is a separate experiment, not a re-measurement of the 7.333-point C-phase primary. It also reproduces the cost of high-rate-source data and again finds no useful learning (Section 4.6).

Contributions. Two complete 2×2 designs separate a dynamic generator policy from a dynamic learner policy, and both show negative paired ACC interactions at the original update rate; a matched learning-rate intervention then compares a high and a low rate in six completed forty-round runs after finite inference-repeatability checks, and all three paired near-probe, ACC and BWT contrasts favour the lower rate without implying that the lower rate forgets nothing (Appendix S and Section 4.2). The centre of the paper is a completed twelve-cell matrix crossing the source of the frozen generated training data with the update rate of the freshly initialised learner, reported with a disclosed resource amendment, two historical rows and all four fixed decoding repeats: its prespecified near-probe screening criterion passes, all twelve held-out training-slice changes are negative, and a confirmation on a new split, new seeds and a new server reproduces the source contrast without finding any positive learning reference (Sections 4.3 and 4.6). We then characterise the six frozen banks themselves and report how three candidate content explanations order against the source effect as measured here (Section 4.4), and we bound the measurement with epoch and multiplicity controls, exact cumulative training exposure and eighteen MATH decoding-budget evaluations, treating the extra-BOS format failure as a separate phenomenon (Appendices S.4, E and G). This is a diagnostic study of retention under answer-verified self-training, not a new self-improvement algorithm. All appendices are supplied as a separate supplementary document; each "Appendix" reference in this paper points to it.

2 Related Work

Parameter-level self-improvement. STaR, ReST, and ReST-EM use generated targets and selection or feedback to bootstrap learning (Zelikman et al., 2022; Gulcehre et al., 2023; Singh et al., 2023). Self-consistency can supply reasoning targets (Huang et al., 2023); Self-Rewarding LMs and CREAM instead construct or regularize self-judged rewards (Yuan et al., 2024; Wang et al., 2024). SEAL learns data and update directives using post-update performance (Zweiger et al., 2025). VSI already evaluates iterative GSM8K self-training under answer-only, majority-vote and reasoning verification with MATH transfer and diversity (Zhang, 2026), so answer-only verification limits are not our novelty claim; we cross generator policy with learner resets over forty small disjoint slices without reproducing VSI. Our self-consistency filter uses majority answers (Wang et al., 2023); that pseudo-label rule is not TTRL's reinforcement-learning reward (Zuo et al., 2025).

Collapse, verification, and specialization. Synthetic feedback can degrade quality or diversity (Shumailov et al., 2023; Alemohammad et al., 2023), and retaining earlier data changes collapse behaviour (Gerstgrasser et al., 2024; Kazdan et al., 2024); that concerns data accumulation, ours parameter accumulation and generator identity. Bertrand et al. (2023) give conditional stability under

accurate initial models and enough clean data; Feng et al. (2024) show imperfect verification can prevent collapse in their settings. Briesch et al. (2023) observe diversity loss despite correct outputs; our answer diversity instead rises during accuracy loss. Format specialization (Wang et al., 2022) and fine-tuning-induced forgetting (Luo et al., 2025) motivate separate retention and compliance diagnostics.

Learning and retention under fine-tuning. FINCH uses a loss-adaptive learning rate to improve the learning–forgetting trade-off; its small-rate baseline also illustrates that reduced forgetting can accompany weaker target-task learning (Prashant et al., 2026). Our matched comparison neither establishes a better trade-off nor reproduces FINCH. What distinguishes the crossing is that freezing the bank unbinds the source of the training data from the rate at which a receiver consumes it, which is what lets a 7.333-point retention difference appear with the receiver rate held fixed – a comparison a loss-adaptive schedule cannot make. Wu et al. (2024b) report a self-improvement reversal in which benchmark gains coexist with narrowing output diversity and weaker out-of-distribution generalisation; our runs have no improvement phase to reverse, and frozen banks cannot narrow in response to a receiver's own updates. Li et al. (2024) attribute tuning-induced forgetting to the sharpness of the loss landscape and mitigate it with sharpness-aware minimisation of the tuning objective, an intervention orthogonal to ours; we measure no loss geometry, so we neither adopt nor exclude that mechanism.

Long-horizon evaluation and continual learning. StreamBench, LifelongAgentBench, and PATH-Bench study extended adaptation through context, memories, or skills with a frozen backbone (Wu et al., 2024a; Zheng et al., 2025; Yang et al., 2026); we instead update weights and add an independent-restart comparison. Our metrics follow Lopez-Paz & Ranzato (2017); Chaudhry et al. (2018); stability denotes behavioral retention under ongoing learning (Mermillod et al., 2013), not a dynamical stability theorem. Less forgetting under online RL than SFT further limits extrapolation from our update recipe (Shenfeld et al., 2025).

Recent direct failure studies. Lin (2026) study rise-and-collapse within code REINFORCE campaigns and compare campaign memory, early stopping and GRPO, making update-rule and budget sensitivity a concern; our oracle-filtered SFT factorial on GSM8K is not that task. Song et al. (2026) compare ungated, external compilation/static-check, and model-coupled review in recursive code fine-tuning, reporting that external filters can slow degradation without eliminating it; static checks differ from our reference-answer filter. We claim no first verification-resistant decline, only what these controls establish in this recipe. KITE changes synthetic-data generation and curation (Luo et al., 2026); RISE combines verifiable-reward reinforcement learning with an extrapolated distillation teacher (Li et al., 2026). Neither is tested here, and both limit claims of inevitable degradation.

3 Experimental Setup

3.1 Self-update loop and controls

A *round* consists of generation, selection, fine-tuning, and evaluation (Figure 1A); round 0 is the base-model evaluation. Let θ_0 be the fixed base checkpoint and ϕ_t the LoRA adapter after round t, with $\phi_0 = \emptyset$. The *stream* is one fixed sequence of $N = 40$ disjoint training/evaluation slices $(\mathcal{T}_t, \mathcal{E}_t)$, presented in the same order to every run. Each round samples $k = 8$ reasoning solutions per training problem at temperature 0.8, selects up to $m = 2$ per problem, and fine-tunes on selected assistant tokens. In **continuous** mode the generator is (θ_0, ϕ_{t-1}) and the learner starts from ϕ_{t-1}; **independent restart** uses θ_0 and an empty adapter each time; **frozen generator** keeps the generator at θ_0 while the learner starts from ϕ_{t-1}; **recursive reset** (`rfresh`, cell R) generates with (θ_0, ϕ_{t-1}) but restarts the learner from an empty adapter each round, a sequential self-distillation chain. Data are newly sampled for each slice; a frozen generator does not mean a fixed training corpus.

3.2 Verification conditions

None keeps random samples without checking their answers. **Parsable** (format check only) requires a finished generation and a parsable final `Answer:` line, without checking correctness. **Self-consistency** (`selfcons`) keeps finished samples agreeing with a strict majority parsed answer (at

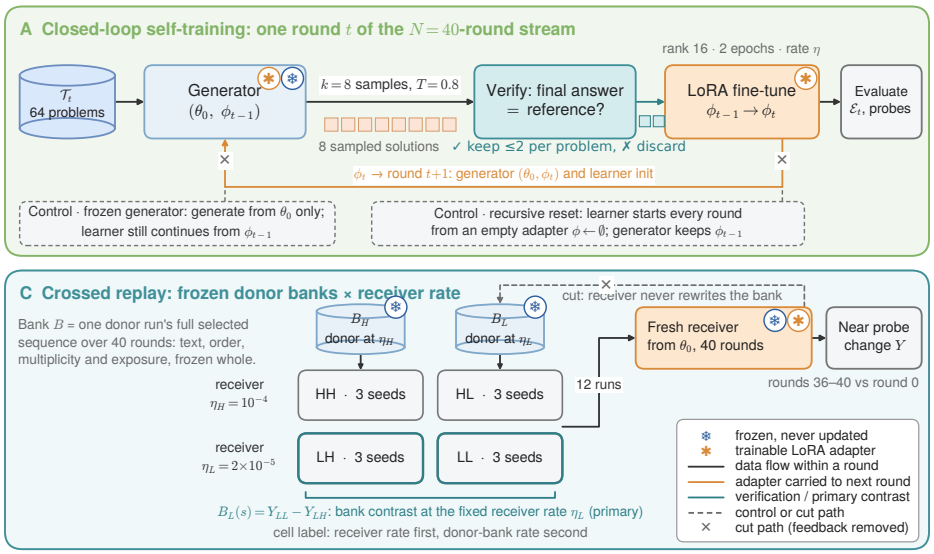


Figure 1: The two designs and what separates them. **A**: one round of the closed loop, with the two controls that hold the generator or the learner static. Because the update rate also fixes the weights that write every later target, a rate difference measured inside this loop is the total effect of the policy, not of the update. **C**: freezing a donor bank cuts that path, so the bank contrast B_L at a fixed receiver rate isolates a data-source component. The transferred treatment is the entire selected sequence, exposure included, not rationale content alone. Panels carry the designs' own names, A and C, so there is no panel B.

least five of eight samples); otherwise the problem contributes nothing. **Oracle** keeps finished samples whose final answer matches the reference; it does not verify reasoning steps. All conditions cap retained samples at two per problem. Selection changes realized counts, lengths, and precision, so exposure is measured rather than assumed identical (Table S16). Only oracle selection accesses reference answers for training-data filtering; held-out and probe items are never training data.

3.3 Benchmarks, fixed evaluation sets, and which runs completed

The primary stream is **GSM8K** (Cobbe et al., 2021), with 64 training and 32 held-out training-split problems per round. Fixed probes comprise 300 GSM8K test items (near), 300 ARC-Challenge test items (far) (Clark et al., 2018), and 100 MATH-500 items (harder OOD math) (Lightman et al., 2023). Numeric and option-letter answers use exact matching; MATH uses loose normalized string matching without symbolic equivalence. The original twelve cells use **Llama-3.1-8B-Instruct** and three seeds each; twelve further oracle-continuous runs (one-epoch, low-rate local, and matched-runtime high/low pairs) are reported separately.

3.4 Metrics

Accuracy and output contract. Prompts require a final `Answer:` <x> line. *Strict* accuracy requires a correct parsed payload on that line; a missing line is wrong. *Lenient* accuracy permits fallback extraction of the last number (GSM8K), answer-like option letter (ARC-C), or boxed expression (MATH). Format compliance measures a parsable answer line regardless of correctness; truncation measures outputs reaching the decoding limit. Strict–lenient differences diagnose recoverable format failures; neither score alone measures latent reasoning ability.

Table 1: What each arm establishes, and what bounds it. It carries no quantities by design: every number it alludes to, with its seed dispersion, is in the section named on the left. Design A is the closed loop, design C the frozen-bank crossing (Figure 1).

Arm	What it establishes	What bounds it
Closed-loop controls and the matched rate intervention, design A (§4.1, §4.2)	Retention follows the recursive update policy rather than the correctness of the retained answers, and a lower rate improves near retention in every seed pair.	The controls move exposure and optimizer scheduling together, and the rate also fixes the weights that write later targets: a total policy effect, not the update alone.
Frozen-bank crossing, design C (§4.3)	At a fixed receiver rate the bank's source still moves near retention; at a fixed bank the rate contrast stays inside the round-zero spread.	The treatment is a whole selected sequence, exposure included, and every receiver loses stream-held-out accuracy.
What the bank treatment is made of (§4.4, §4.5)	Measured bank summaries do not identify a content mechanism; authored rationales decline too, and matching presented tokens or optimizer steps against them does not remove it.	An ordering is descriptive: it names no acting property. The matching is against that reference, not inside the crossing, so no rate contrast at a fixed token dose exists.
Replications: second receiver family, split site (Appendix N, §4.6)	The source contrast reappears in another pre-training family under both parsers, and survives a new partition, new seeds and a new server on pre-fixed endpoints.	The second family's round-zero evaluations spread wider and miss a prespecified agreement check; the confirmation inherits some donor banks, analysed them after the cutoff, and its learning reference is negative.

Continual-learning summaries. Let $R_{t,i}$ be strict accuracy on held-out slice $\mathcal{E}_i$ after update t, with $R_{0,i}$ the base-model score. We use

$$\mathrm{ACC} = \frac{1}{N}\sum_{i=1}^{N} R_{N,i}, \qquad \mathrm{BWT} = \frac{1}{N-1}\sum_{i=1}^{N-1}(R_{N,i} - R_{i,i}),$$

$$\mathrm{FWT} = \frac{1}{N-1}\sum_{i=2}^{N}(R_{i-1,i} - R_{0,i}), \qquad \mathrm{FGT} = \frac{1}{N-1}\sum_{i=1}^{N-1}\left(\max_{i\le t\le N} R_{t,i} - R_{N,i}\right).$$

FWT evaluates incoming slice i after update $i-1$, before training on it. ACC concerns stream-held-out slices, not the fixed near probe. BWT and FGT exclude the final slice. Distinct-answer count among four samples measures answer diversity, not rationale diversity.

The **sustained-degradation onset** is the first $t \le N-2$ for which near strict accuracy satisfies $p_{t'} < p_0 - 0.05$ for every $t' \in \{t, t+1, t+2\}$. This criterion was fixed before the v2 runs (2026-09-05). Onset counts seeds meeting this criterion, not catastrophic collapse; a seed may subsequently recover. The primary probe effect for seed s is $\Delta_s = \frac{1}{5}\sum_{t=36}^{40} p_{s,t} - p_{s,0}$; we report its mean and sample standard deviation across three seeds, together with every seed of the new controls. Round-40 endpoints are secondary for these probe analyses; the separate authored-rationale reference uses full-stream round-40 minus own-round-zero change as primary (Section 4.5). CC separately prespecifies 960-item stream learning and final-five-round near retention. All policy contrasts are descriptive and paired by seed on a shared stream. No pooled-label permutation probabilities, population significance, or equivalence claims are made. Identical base weights yield near scores spanning 82.67–85.00% across the 36 round-zero evaluations, a strict near spread of 2.333 points; base-model MATH scores span 34–39%. The offline receivers of Section 4.3 instead all record the same round-zero near score, 86.000 percent strict and 86.333 percent lenient: they are evaluated under the separately frozen sparse schedule with batch-invariant inference, and within each seed all four cells reuse one byte-identical round-zero record. The shared record has zero variance by construction and does not estimate repeat-evaluation variability. The historical 2.333-point spread is a descriptive reference from another setup; uncertainty under the crossing's setup has not been separately quantified. The second receiver family evaluates round zero in every cell and has an observed 3.000-point spread. This variability is not separated from training randomness; stronger inference needs more data orders and controlled repeat evaluations.

4 RESULTS

4.1 WHAT THE ORIGINAL GENERATOR AND LEARNER CONTROLS ESTABLISHED

The original oracle continuous condition has a near change of -5.67 ± 1.01 points despite perfect retained-answer precision, and both the oracle and unverified generator/learner factorials show negative paired ACC interactions in all three seeds; these compare dynamic policies, not fixed rationale distributions. Full tables, trajectories, per-seed interactions and the retained control, format and diversity observations are in Appendices S and I; MATH remains budget- and scorer-dependent, the coarse sustained-degradation criterion is insensitive to several observed declines, and the later source comparison removes neither qualification.

4.2 A COMPLETED MATCHED LEARNING-RATE INTERVENTION

We compare two epochs at 10^{-4} and 2×10^{-5}, retaining up to two oracle-accepted samples per problem. Seed pairs share model bytes, data order, and software. All six runs complete forty rounds with identical paired round-zero evaluations. Appendix K reports the frozen runtime-admission checks, earlier failures, and finite scope of repeatability.

The trajectories appear in Appendix Figure S6. Lowering the rate improves the primary near change by $+7.84 \pm 1.65$ points, with paired differences $+6.47/+9.67/+7.40$. ACC improves by $+7.84 \pm 0.36$ and BWT by $+4.86 \pm 0.23$ points. Near losses are small but nonzero in two low-rate seeds, and BWT remains negative in all three. Item-level alignment also shows the low-rate near change combines gains of 6.133 points with losses of 6.289 points over the final five rounds and three seeds (Appendix M), so near-zero net change coexists with substantial turnover of correct answers. Far and operational MATH contrasts also favor the lower rate (Table S10); they are secondary fixed-probe measurements, not independent task replications.

4.3 CROSSING THE SOURCE OF THE FROZEN GENERATED TRAINING DATA WITH THE UPDATE RATE OF A FRESHLY INITIALISED LEARNER

Because the matched recursive-rate comparison changes both updates and later training data, we ran a separate exploratory offline $2 \times 2 \times 3$ comparison that held each donor bank fixed while varying the receiver rate. A bank is a donor's full selected sequence, treated jointly: text, coverage, multiplicity, order, exposure. Every receiver started from the base checkpoint and trained forty rounds without regenerating its bank. We use HH, HL, LH, LL with receiver rate first and donor-bank rate second (Figure 1C); both rates and the near endpoint are unchanged from the frozen design. Appendix T defines the five contrasts.

Primary membership and amendment. The primary matrix combines four prospectively specified replacement seed-zero runs with eight original seed-one/two runs. The resource amendment followed partial result visibility and changed the original device/no-rerun commitment. Original seed-zero HH/LL runs remain disclosed, excluded by the prospective membership rule rather than their outcomes. There are twelve primary cells, fourteen total trained runs, and four decoding repeats. Substituting the excluded historical seed-zero LL row, whose strict near change is -1.333 points, for its replacement gives $B_L(s_0) = +7.000$ and a mean of 7.267 instead of 7.333 points, so the amendment does not decide the screened outcome: the criterion passes either way. Appendix V gives the device constraint, chronology, historical endpoints, and accounting; Table 2 preserves within-seed device grouping.

A data-source contrast at the fixed low update rate. Appendix Figure S9 visualizes the complete contrasts. The strict primary $B_L = Y_{L,L} - Y_{L,H}$ is 7.200, 8.333 and 6.467 points for seeds zero, one and two: mean 7.333 and sample SD 0.940 points. It meets the criterion fixed in advance of these runs — a mean of at least three points with a positive value in each seed — which we call the near-probe screen. Lenient B_L is $7.333/8.200/6.467$, with the same mean 7.333 but a different sample SD of 0.867. These are three stochastic runs on the same question stream, not independent data-set replications; the SD is descriptive rather than a confidence interval.

Table 2: The complete C primary matrix: Y is the mean near accuracy over rounds 36–40 minus the same run's round 0, strict and lenient, in percentage points. Every receiver starts from the same base checkpoint, whose round-zero near score is 86.000 percent strict and 86.333 percent lenient, so within each parser $\overline{p}_{36:40} = p_0 + Y$. Branch C holds the four new seed-zero cells and Original the eight seed-one/two cells; device labels are anonymous and preserve within-seed grouping. The excluded historical seed-zero diagonal is disclosed separately, not averaged in.

		HH		HL		LH		LL	
Seed	Branch / device	strict	lenient	strict	lenient	strict	lenient	strict	lenient
0	C / D	-7.333	-7.600	-2.267	-2.333	-8.333	-8.400	-1.133	-1.067
1	Original / B	-10.133	-10.267	-0.600	-0.933	-8.200	-8.400	+0.133	-0.200
2	Original / C	-7.400	-7.733	-1.667	-1.867	-7.000	-7.200	-0.533	-0.733

Table 3: All five near-strict contrasts (percentage points). B_L is the primary screen and B_H the same bank contrast at the high receiver rate; A_H and A_L lower the receiver rate with the bank fixed to H or L; $I = B_H - B_L$. All three seeds and sample SD are shown; secondary contrasts do not change the primary criterion.

Contrast	Seed 0	Seed 1	Seed 2	Mean	Sample SD
B_L	+7.200	+8.333	+6.467	+7.333	0.940
B_H	+5.067	+9.533	+5.733	+6.778	2.410
A_H	-1.000	+1.933	+0.400	+0.444	1.467
A_L	+1.133	+0.733	+1.133	+1.000	0.231
I	-2.133	+1.200	-0.733	-0.556	1.674

The low-rate receiver's strict near changes are $-8.333/-8.200/-7.000$ with the high-rate bank and $-1.133/+0.133/-0.533$ with the low-rate bank. Their means are -7.844 and -0.511 points. The lower receiver rate therefore did not preserve near performance for both tested sources. This concerns scored responses under this decoder and recipe, not capability damage.

The crossing does not isolate a content mechanism. The high-rate bank contrast is also positive in every seed (B_H mean 6.778 points), whereas the fixed-H-bank receiver-rate contrast A_H has mean 0.444 and includes a negative seed. The interaction I has mean -0.556 and mixed signs (Table 3). Their sample SDs are 1.467 for A_H, 0.231 for A_L and 1.674 for I, and all three means lie inside the 2.333-point spread of the thirty-six round-zero evaluations of identical base weights (Section 3.4), whereas B_L and B_H are about three times that spread. The historical spread uses a different evaluation setup and is only a descriptive reference; measurement uncertainty for this crossing remains unquantified. These three-seed contrasts establish neither a zero rate effect nor a mediated fraction of the closed-loop difference; the transferred treatment is the whole frozen bank, so rationale quality, token dose and optimizer-step count are not separated. Near-format B_L is $+0.667$ points and non-stop-termination B_L is -0.867 points.

Retention differences in the crossed experiment without positive held-out learning. A separate descriptive recount of the complete 1,280 stream-held-out items at round 40 versus round 0 gives a negative strict change for *every one of the twelve primary cells*, ranging from -8.984 to -0.547 points. Appendix W reports all cells, including the historical rows separately, and Appendix R breaks the same endpoint down by item. This post-result descriptive endpoint is not an added preregistered gate and does not replace the near screen. It provides no evidence of useful self-improvement over the base in these runs.

The source contrast, but not the self-improvement, survives a second receiver family. Every receiver above is an instance of the model that wrote the banks, so we re-ran the same four frozen banks at seeds zero and one in front of Qwen2.5-7B-Instruct, everything else unchanged (eight cells; Appendix N). Strict B'_L is $+4.867$ and $+4.800$ points, mean 4.833, and $+4.800/+5.133$ under the lenient parser, against 7.333 here; the banks remain Llama-generated, so this varies the receiver and not the donor, and this arm's own round-zero spread is 3.000 points. All eight of its cells decline

under the lenient parser, whose round-zero score already credits the answer-contract compliance this receiver acquires while training, so the arm adds no evidence of useful self-improvement either.

4.4 What distinguishes the contents of the two banks

The source contrast is an effect of an object the paper has so far described only by counts, so we characterised the six banks from the stored bank files (Table S19; definitions, length distributions and the per-item join in Appendix O). We compare three descriptive summaries across seeds; their ordering is not a causal test. Supervision volume: supervised tokens per example span 205.66–214.00 in the high-rate banks and 200.46–207.79 in the low-rate banks, and the ordering reverses in seed one, whose source contrast is the largest of the three ($+8.333$ points): its low-rate bank holds both more supervised tokens ($1{,}035{,}820$ against $1{,}025{,}000$) and longer targets, so a simple explanation in which fewer supervised tokens consistently improve retention does not fit all three seeds. Lexical variety: distinct-4-gram ratios span 0.4434–0.4720 against 0.4628–0.4713, again without a consistent sign, and mean final-answer digit counts lie between 2.312 and 2.318 in all six banks. Lexical proximity is higher in the more damaging banks: the high-rate banks cover 18.04–18.26 percent of the near probe's 3-gram types and the low-rate banks 17.50–17.59 percent, so in every seed the more damaging bank is the one closer to the probe; per item that coverage difference has weak measured correlation with the item-level source effect $d_{i,s}$, with largest absolute Pearson correlation 0.111 and largest absolute Spearman correlation 0.045 over 300 probe items.

Two measured properties do order with the effect in all three seeds. The low-rate banks are close to three times as verbatim-redundant, 6.42–6.60 against 2.14–2.41 percent of rows repeating the normalised text of another row, and in all six banks every such repeat is one question sampled twice rather than two different questions; their solutions also spread fewer words over more lines, 10.95–11.72 lines and 137.87–140.91 words against 9.79–10.85 lines and 141.55–151.97 words. We manipulated neither property, and with three banks per source a co-varying property cannot be separated from any other that orders the same way, so these are characterisation, not mechanism. The summaries constrain simple monotonic accounts but neither identify nor exclude causal contributions of supervision volume, n-gram diversity or lexical proximity. Nonlinear effects and interactions remain possible because these properties were not independently manipulated.

4.5 An exploratory authored-rationale learning reference

After observing C we evaluated $C_{L\mathrm{gold}}$, an authored-rationale reference under the low-rate schedule: the same records with each generated solution replaced by the dataset's authored one. Its primary outcome, round-40 minus own-round-zero accuracy on the full 1,280 stream-held-out items, is negative in every seed under both parsers (mean -11.328 strict, -11.458 lenient); counting every missing or length-terminated endpoint as correct leaves all seeds negative (Appendices P and P.3). That arm changes provenance, supervision dose and target length at once: the authored corpus carries $1{,}183{,}988$ supervised tokens against $3{,}065{,}605$ in the corresponding low-rate banks, 2.59 times less supervision over the same records at the same $3{,}824$ steps, and mean endpoint output length falls from $206.323/205.670/205.609$ to $79.058/80.291/76.612$ tokens. Two banks built from B_L without new authored text separate these: D_1, matched to the authored token budget with rationales whole, reaches $+1.556$ points, positive in every seed; D_2, matched to that budget and to each target's length by truncating the generated rationale, reaches -11.333, past the reference's own -10.378, not near the low-rate receiver's -0.511. Cutting supervision to the authored budget does not reproduce the decline; matching the authored lengths does, so the decline needs no authored targets. Truncation shortens reasoning as well as targets. Six anchor cells re-measure both reference points in the batch that ran these banks and reproduce them inside the pre-registered band, putting D_2's explained fraction at 1.160 there and 1.097 on the earlier batch; Appendices P.1 and P.2 give per-seed values, the non-unanimous seed, the amended round-zero check and every cell's round zero.

4.6 A completed confirmation on a new data split, new seeds, and a new server

The crossing above used one partition, seeds $0, 1, 2$ and one set of devices; we repeated it on a predetermined disjoint partition with seeds $101, 102, 103$ on a different server. Per seed, a high-rate and a low-rate donor run produce two frozen banks; two fresh learners train at the fixed low update rate,

one on each bank; a third trains on the same records with the dataset's authored solutions instead, preserving question identities, multiplicities and order. All fifteen forty-round jobs completed.

The data-source contrast reproduces off the original split. At the fixed low update rate the near-domain strict advantage of low-rate-source over high-rate-source data is $B'_L = +4.40/+7.67/+5.47$ points for seeds 101, 102 and 103: mean $+5.84$, sample SD 1.67. Lenient values are positive in every seed too (mean $+5.82$), so the criterion fixed in advance — a strict mean of at least three points, every strict seed above zero, every lenient seed above zero — passes, and the narrow transfer estimate is recorded as confirmed. These are two experiments, not two measurements of one quantity: the 7.333-point contrast quoted above and in the abstract is the C-phase primary on the original split with seeds $0/1/2$, whereas $+5.84$ points is the split-site confirmation on the new split with seeds $101/102/103$. They agree in direction; the confirmation is the smaller, and the change of split and seeds is part of why.

High-rate-source data damages; low-rate-source data does not restore. The receiver given high-rate-source data loses near-domain strict accuracy in every seed ($-5.13/-5.60/-4.33$, mean -5.02, sample SD 0.64), so the prespecified absolute-damage criterion, a mean of at most -3 points with every seed negative, passes too. Its low-rate-source counterpart is near flat on the same probe (mean $+0.82$, sample SD 1.43). Flat is not learning: the prespecified usefulness condition for self-training, requiring a mean full-stream gain of at least three points, is false at both source conditions.

The authored-rationale reference declines again, so the diagnostic claim fails. The third learner of each seed declines on both endpoints in every seed: mean -11.42 points (sample SD 1.12) on the 960 held-out items and $-12.27/-12.33/-12.60$ on the near probe (mean -12.40, sample SD 0.18). These are receiver-side quantities for the authored-rationale condition, not donor quantities. Because that reference does not satisfy the stream-learning and near-retention criteria together, the prespecified diagnostic claim fails at its gate and no secondary contrast rescues it. What transfers is a source-attributable retention difference, not a useful learning–retention tradeoff against supervised text; the latter is contradicted at both rates and in every seed. Per-job accounting, remaining gate values, the inherited-donor boundary and the post-cutoff analysis timing are in Appendix Q.

Second training stream. The same crossing on ARC-Challenge completes eight receivers. Its primary source contrasts are $+1.133$ and -2.400 points across two seeds, both within the 3.333-point round-zero spread. This arm does not reproduce the GSM8K source contrast at its measurement resolution (Appendix X).

5 Discussion and Limitations

Limitations. We test one donor, two receiver families and fixed GSM8K/ARC-Challenge orders. The GSM8K crossing has three seeds; the second-receiver and ARC-Challenge arms have two. The ARC-Challenge arm does not establish cross-stream replication (Appendix X). Seeds do not sample independent task distributions. The original factorial only tests the high-rate two-sample recipe. The low-rate result neither proves equivalence to zero loss nor extends beyond forty rounds. Other donors, task orders, verifiers and reset/regeneration tradeoffs remain untested. The second family's eight round-zero evaluations of identical weights span 3.000 points, failing the prespecified 0.34-point agreement check; its baseline is noisier. We make no direct VSI, FINCH, or KITE comparison. Further intervention, evaluation and protocol limitations appear in Appendix Y.

Conclusion. Correct-answer filtering alone does not determine retention. The matched recursive comparison favors the lower-rate policy but also changes subsequent data. Crossing frozen banks with receiver rates gives a 7.333-point source contrast at the low receiver rate; confirmation with a new partition, seeds and server gives $+5.84$ points. This concerns whole-bank assignment, not content, mediation or capability damage. Every C receiver, the authored-rationale reference and the confirmation reference lose stream-held-out accuracy. The source contrast survives changes of split, seeds, machine and receiver family but does not reproduce on ARC-Challenge at the tested budget. Neither GSM8K rate yields useful self-improvement.

AI USE STATEMENT

In this work, we used generative AI tools (Claude Code with Opus-5.5) for cleaning and reformatting data, including stream construction from GSM8K, MATH-500, and ARC-C and the answer-extraction rules.

ETHICS STATEMENT

We use open-weight models and public benchmarks, without human subjects or private data.

REPRODUCIBILITY STATEMENT

Local reproducibility materials include raw per-round evaluation and training records, fixed split identifiers, configurations, and a standalone CPU audit with input hashes. Every main trajectory is regenerated from these records, and no original main run is removed. The audit explicitly distinguishes retained-corpus token counts from tokens presented across epochs. Environment pins and raw provenance for the completed studies are retained internally. The first CC attempt has an incomplete report with unevaluable effects; the second attempt supplied complete scientific evidence, including a guardian exit code of zero, per-job ledgers and raw verification for the nine jobs run on the confirmation server, and a pinned source-host audit certificate for the six inherited donor banks. Gold tables are generated from the complete recorded CPU readout; independent audits reproduced the raw-score quantities and the bounded output-contract diagnostic. Public release packaging and independent end-to-end reproduction remain pending. The appendices are provided as a separate supplementary PDF.

REFERENCES

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850*, 2023.

Quentin Bertrand, Avishek Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. *arXiv preprint arXiv:2310.00429*, 2023. URL https://arxiv.org/abs/2310.00429.

Martin Briesch, Dominik Sobania, and Franz Rothlauf. Large language models suffer from their own output: An analysis of the self-consuming training loop. *arXiv preprint arXiv:2311.16822*, 2023. URL https://arxiv.org/abs/2311.16822.

Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. arXiv:1801.10112.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yunzhen Feng, Elvis Dohmatob, Pu Yang, François Charton, and Julia Kempe. Beyond model collapse: Scaling up with synthesized data requires verification. *arXiv preprint arXiv:2406.07515*, 2024. URL https://arxiv.org/abs/2406.07515.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*, 2024.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (ReST) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023. URL https://arxiv.org/abs/2308.08998.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.67. URL https://arxiv.org/abs/2210.11610.

Joshua Kazdan, Rylan Schaeffer, Apratim Dey, Matthias Gerstgrasser, Rafael Rafailov, David L. Donoho, and Sanmi Koyejo. Collapse or thrive? perils and promises of synthetic data in a self-generating world. *arXiv preprint arXiv:2410.16713*, 2024.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4297–4308, 2024. doi: 10.18653/v1/2024.findings-emnlp.249. URL https://aclanthology.org/2024.findings-emnlp.249/.

Yang Li, Semih Yavuz, and Shafiq Joty. Rise: Recursive improvement via self-extrapolating policy distillation. *arXiv preprint arXiv:2609.05295*, 2026. URL https://arxiv.org/abs/2609.05295.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Jianzhe Lin. Self-improvement can self-regress: The rise-and-collapse failure mode of LLM self-training. *arXiv preprint arXiv:2606.21090*, 2026. URL https://arxiv.org/abs/2606.21090.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems 30*, 2017. arXiv:1706.08840.

Xiaonan Luo, Yue Huang, Kehan Guo, Ping He, Chuan Zou, Ting Hua, and Xiangliang Zhang. Learning from synthetic data without model collapse in iterative instruction tuning. *arXiv preprint arXiv:2607.17043*, 2026. URL https://arxiv.org/abs/2607.17043.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025. doi: 10.1109/TASLPRO.2025.3606231. URL https://arxiv.org/abs/2308.08747.

Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in Psychology*, 4:504, 2013. doi: 10.3389/fpsyg.2013.00504.

Parjanya Prajakta Prashant, Jiongli Zhu, Aldan Creo, and Babak Salimi. Fine-tuning without forgetting via loss-adaptive learning rates. *arXiv preprint arXiv:2605.20005*, 2026. URL https://arxiv.org/abs/2605.20005v1.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. Rl's razor: Why online reinforcement learning forgets less. *arXiv preprint arXiv:2509.04259*, 2025.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023. URL https://arxiv.org/abs/2312.06585. Accepted to Transactions on Machine Learning Research.

Xinyuan Song, Zekun Cai, and Liang Zhao. When AI reviews its own code: Recursive self-training collapse in code LLMs. *arXiv preprint arXiv:2606.28438*, 2026. URL https://arxiv.org/abs/2606.28438.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. arXiv:2203.11171 (2022); published at ICLR 2023.

Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit S Dhillon, and Sanjiv Kumar. Two-stage LLM fine-tuning with less specialization and more generalization. *arXiv preprint arXiv:2211.00635*, 2022. URL https://arxiv.org/abs/2211.00635.

Zhaoyang Wang, Weilei He, Zhiyuan Liang, Xuchao Zhang, Chetan Bansal, Ying Wei, Weitong Zhang, and Huaxiu Yao. CREAM: Consistency regularized self-rewarding language models. *arXiv preprint arXiv:2410.12735*, 2024. URL https://arxiv.org/abs/2410.12735.

Cheng-Kuang Wu, Zhi Rui Tam, Chieh-Yen Lin, Yun-Nung Chen, and Hung-yi Lee. Streambench: Towards benchmarking continuous improvement of language agents. In *Advances in Neural Information Processing Systems 37, Datasets and Benchmarks Track*, 2024a. arXiv:2406.08747.

Ting Wu, Xuefeng Li, and Pengfei Liu. Progress or regress? self-improvement reversal in post-training. *arXiv preprint arXiv:2407.05013*, 2024b. URL https://arxiv.org/abs/2407.05013.

Xidong Yang, Xingyi Zhang, Wenhao Li, Wenyan Liu, Junjie Sheng, Yun Hua, Wei Yin, Tao Fang, Chuyun Shen, and Xiangfeng Wang. Path-bench: Path-dependent evaluation of lifelong agents. *arXiv preprint arXiv:2608.01149*, 2026.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. arXiv:2401.10020.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems 35*, 2022. URL https://arxiv.org/abs/2203.14465.

Xinyu Zhang. Reliable self-improvement training by verifying reasoning, not just answers. In *ICLR 2026 Workshop on LLM Reasoning*, 2026. arXiv:2603.21558.

Junhao Zheng, Xidi Cai, Qiuke Li, Duzhen Zhang, ZhongZhi Li, Yingying Zhang, Le Song, and Qianli Ma. Lifelongagentbench: Evaluating llm agents as lifelong learners. *arXiv preprint arXiv:2505.11942*, 2025.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. Ttrl: Test-time reinforcement learning. In *Advances in Neural Information Processing Systems 38*, 2025. arXiv:2504.16084.

Adam Zweiger, Jyothish Pari, Han Guo, Yoon Kim, and Pulkit Agrawal. Self-adapting language models. In *Advances in Neural Information Processing Systems 38*, 2025. arXiv:2506.10943.