Model Collapse in Multi-Model Feedback Loops Is Locally Reversible in Weight Space

Anonymous ACL submission

Abstract

Model collapse has primarily been studied in single-model feedback loops, yet future training corpora will more likely contain generations from many different models. We study *multi-model* collapse: at each generation, distinct model families generate synthetic text, their outputs are pooled and shuffled, and every model trains on the shared synthetic corpus. Across full-parameter 1B–1.5B loops and QLoRA-trained 7B–8B loops, multi-model pooling still produces collapse. We then ask whether the collapse trajectory is locally reversible in parameter space. Given a current checkpoint θ_i and a one-step-further collapsed checkpoint θ_{i+1}, we form $\hat{\theta}_{i-1} = \theta_i - \alpha(\theta_{i+1} - \theta_i)$ and compare it with both θ_i and the earlier target θ_{i-1}. This task-vector negation reduces held-out corpus perplexity across all families and closes significant downstream gap to the target checkpoint without using any external data. Trajectory analysis shows that adjacent collapse steps are positively aligned but not colinear, and that their norms shrink across generations. Local cosine alignment explains why one-step reversal works, while non-colinearity explains why two-step extrapolation degrades. These results suggest that, at least locally, multi-model collapse behaves more like directional drift in weight space than like progressive destruction of capability.

1 Introduction

Language models increasingly train on text produced by other language models. As generated text spreads across the web, future training corpora will contain a growing share of model outputs, and successive models will retrain on this mixture. This feedback can produce *model collapse*: a progressive loss of distributional coverage, diversity, and quality under recursive training on self-generated data (Shumailov et al., 2024; Alemohammad et al., 2023).

Most empirical studies of language-model collapse consider a single-model feedback loop, in which one model generates synthetic text, retrains on that text, and repeats this process across generations. Although prior work varies the training mixture, filtering procedure, and data-generation strategy, the synthetic text is almost always produced by a single model family (Shumailov et al., 2024; Dohmatob et al., 2024; Gerstgrasser et al., 2024; Zhu et al., 2025; Kazdan et al., 2025; Feng et al., 2025). Wang et al. (2025) extend the analysis to a network of LLMs interacting through a shared retrieval database, but the constituent models remain frozen. Collapse is measured at the level of output distributions rather than model parameters. The assumption of frozen or single-source generation is unlikely to match future web-scale corpora. As generated text becomes widespread online, training data will contain outputs from many model families and providers (Dolezal et al., 2026). The more realistic feedback loop is a collective and recursive setting where models contributes to trains on a shared synthetic pool.

We study this multi-model collapse setting. At each generation, multiple open-weight models generate synthetic text from a common prompt set. Their generations are pooled, shuffled, and used to train all model in the next generation.

Beyond measuring collapse, we ask whether the observed degradation can be partially reversed without using external data. Given a current checkpoint θ_i and the next checkpoint θ_{i+1} produced by one more round of synthetic-data training, we treat their difference $\theta_{i+1} - \theta_i$ as an observed collapse update. We then subtract a scaled version of this update from the current checkpoint,

$$\hat{\theta}_{i-1} = \theta_i - \alpha\,(\theta_{i+1} - \theta_i), \qquad (1)$$

producing a one-step rewind in weight space. This procedure uses only model checkpoints from the collapse loop and requires no additional training

data. We evaluate whether the rewound checkpoint recovers, or improves upon, the perplexity and task accuracy of the pre-collapse model.

Contributions.

- We introduce a multi-model collapse setting for language models, in which multiple distinct models iteratively generate synthetic data, pool their generations, and retrain on the resulting shared corpus.
- We show that multi-model pooling still results in model collapse, indicating that model diversity alone is insufficient to prevent collapse.
- We show that subtracting a scaled collapse update in weight space, using only checkpoints from the collapse loop, repairs model performance. Rewound checkpoints reduce perplexity and recover downstream accuracy relative to the collapsed checkpoint without requiring additional data.
- We analyze the geometry of collapse updates: norms shrink across generations and adjacent updates have consistently positive cosine similarity, explaining why one-step reversal succeeds.

2 Related Work

2.1 Model Collapse

Model collapse refers to the progressive loss of distributional coverage, diversity, and quality that occurs when generative models are recursively trained on their own outputs (Shumailov et al., 2024). The phenomenon has been documented across a wide range of generative families, including variational autoencoders and Gaussian mixture models (Shumailov et al., 2024), image diffusion models (Alemohammad et al., 2023; Hu et al., 2025), and molecule conformation models (Gerstgrasser et al., 2024), suggesting that recursive self-training is a general failure mode of generative modeling.

For language models, most empirical work studies a single model retraining on its own outputs. The closest multi-source setup to ours is Wang et al. (2025), which simulates a network of LLMs generating and sharing through a common retrieval database. That setup keeps the constituent models frozen and tracks collapse at the level of output distributions; we instead let multiple transformer LLMs each retrain on a shared pool of one another's generations and measure collapse at the level of model parameters.

Collapse has been attributed to sampling bias against tail events, approximation error accumulating across generations, and the gradual narrowing of the support of the generative distribution (Shumailov et al., 2024). We show that despite these mechanisms, the collapse trajectory is locally structured enough that a future collapse step can be used to estimate a prior one.

2.2 Mitigation Strategies

Existing mitigations for model collapse largely intervene at the data level. A common approach is real-data anchoring: Gerstgrasser et al. (2024) show that retaining real data while accumulating synthetic data can avoid collapse across language models, variational autoencoders, and diffusion settings. However, Dohmatob et al. (2025) show theoretically, in a high-dimensional regression setting, that collapse persists under naive real–synthetic mixing whenever the synthetic fraction does not vanish, suggesting that anchoring alone is not sufficient when synthetic data is abundant.

A second line of work targets the synthetic data itself, through training-mixture design, generated-text detection, or output verification (Drayson et al., 2025; Kazdan et al., 2025; Feng et al., 2025; Yi et al., 2026). Zhu et al. (2025) propose token-level editing of synthetic samples to preserve distributional tails in language model training, while Alemohammad et al. (2024) use self-generated synthetic images as negative samples to improve image generation quality.

Our intervention targets a different avenue. We use adjacent checkpoints from the collapse trajectory to move the current model in the opposite parameter-space direction.

2.3 Task Arithmetic and Negative Extrapolation

Task arithmetic treats differences between model checkpoints as semantically meaningful directions in parameter space (Ilharco et al., 2023). Adding such task vectors composes capabilities, while negating them removes or suppresses behaviors, providing a lightweight alternative to retraining. Subsequent work has extended this paradigm with merging procedures that resolve interference between task vectors (Yadav et al., 2023), tangent-space formulations that improve compositionality

(Ortiz-Jimenez et al., 2023), and applications to machine unlearning (Kim et al., 2025; Dong et al., 2025).

The most closely related prior work is Neon (Alemohammad et al., 2026), which applies negative extrapolation to mitigate collapse in generative image models. Given a reference θ_r and a self-trained θ_s, Neon forms $(1+w)\theta_r - w\,\theta_s$, which is algebraically equivalent to our Eq. 4. However, our work has different conceptual setting. Neon uses negative extrapolation from a healthy reference model that has been mildly self-trained, recovering a slightly degraded image generator. We apply the same operation along a long multi-model collapse trajectory in language models, where no healthy reference exists. That the operation still identifies a useful repair direction many generations into collapse is not predicted by Neon's single-step image-domain results.

3 Method

Multi-Model Recursive Synthetic Training Let M be a set of models $m \in M$ placed in an iterative loop. Let θ_0^m be an initial language model fine-tuned on human-written text. At generation i, model θ_i^m generates a synthetic corpus D_i^m. The corpora are concatenated and shuffled,

$$D_i^{\text{mix}} = \text{shuffle}\left(\bigcup_{m=1}^{M} D_i^{(m)}\right), \tag{2}$$

and every family is continued-pretrained on the same mixed corpus to obtain $\theta_{i+1}^{(m)}$ from $\theta_i^{(m)}$.

Task Vector Negation We use an index convention centered on the model being repaired. Here θ_i denotes the *current* checkpoint, θ_{i-1} denotes the desired earlier target, and θ_{i+1} denotes the checkpoint obtained by taking one additional recursive synthetic-data training step from θ_i. We define the locally observed collapse step from the current checkpoint as

$$\tau_{i+1} = \theta_{i+1} - \theta_i. \tag{3}$$

We then estimate the previous checkpoint using task vector negation (Ilharco et al., 2023).

$$\hat{\theta}_{i-1} = \theta_i - \alpha\tau_{i+1}. \tag{4}$$

Intuitively, $\hat{\theta}_{i-1}$ estimates an earlier, healthier checkpoint by moving against the most recent observed collapse direction. We evaluate several values of α ranging from 0.25 to 2. We also evaluate a closed-form two-step extrapolation, derived in Appendix D,

$$\hat{\theta}_{i-2} = \theta_i - (\alpha + \alpha^2)\tau_{i+1}. \tag{5}$$

This asks whether the single observed future collapse direction remains useful when extrapolated two reverse steps. If the collapse direction were globally linear, this extrapolation would recover θ_{i-2} comparably well to a one-step estimator that observes θ_{i-1} and θ_i.

4 Experimental Setup

4.1 Data and Evaluation

All experiments use C4 English for human-generated training data. Both 1B–1.5B and 8B multi-model collapse experiments use 200k training examples, 10k validation examples, and 10k test examples, with maximum sequence length 512. This corresponds to 102.4M training tokens. For synthetic data generation, each model produces 20k continuations per generation from 64-token C4 training-split prefixes, using temperature 0.7, top-p 0.9, and repetition penalty 1.2.

We report perplexity on held-out real C4 test data and downstream normalized accuracy using `lm-evaluation-harness` (Gao et al., 2024). The downstream suite includes ARC-Challenge, ARC-Easy, BoolQ, HellaSwag, OpenBookQA, PIQA, SciQ, TruthfulQA-MC2, and WinoGrande. We report average normalized accuracy across these tasks for each checkpoint.

4.2 Model Settings

Full-rank 1B–1.5B. The multi-model setting uses Llama 3.2 1B (Grattafiori et al., 2024), Gemma 3 1B (Team et al., 2025), and Qwen 2.5 1.5B (Qwen et al., 2025) base models. All undergo full-parameter continued pretraining for one epoch per generation, learning rate 2×10^{-5}, cosine schedule, batch size 20, and sequence length 512.

QLoRA 7B–8B. The larger multi-model setting uses Llama 3.1 8B (Grattafiori et al., 2024), OLMo 3 7B (Olmo et al., 2026), and Qwen3 8B (Yang et al., 2025) base models. Models are continued-pretrained with QLoRA for one epoch per generation, learning rate 10^{-4}, effective batch size 32, LoRA rank 64, LoRA alpha 128, dropout 0.05, and all-linear target modules.

Family	Gen0	Gen3	Gen6	Gen7	Gen9	Gen10
Llama	13.67	80.33	161.98	188.20	242.57	293.09
OLMo	18.03	44.63	78.79	89.41	106.61	119.56
Qwen	16.85	34.77	59.49	66.86	83.46	99.92

Table 1: Held-out C4 perplexity during 7B–8B multi-model recursive training.

Family	Gen0	Gen3	Gen5	Gen7	Gen9
Llama	19.27	102.96	250.94	431.18	682.15
Gemma	20.53	88.65	182.04	271.79	365.45
Qwen	21.08	75.30	166.72	274.94	411.40

Table 2: Held-out C4 perplexity during full-parameter 1B–1.5B multi-model recursive training.

5 Results

We organize the results around multi-model synthetic feedback. We first show that shared-pool recursive training produces model collapse in both full-parameter 1B–1.5B models and QLoRA-trained 7B–8B models. We then show that local task-vector negation reduces held-out perplexity and improves downstream task scores in both collapse scenarios.

The analogous single-model collapse and recovery results are provided in Appendix A as a baseline sanity check. A single model recursively trained on its own outputs collapses as in prior work, and one-step task-vector negation reduces its perplexity. We treat the multi-model pooled setting as our central regime.

5.1 Multi-Model Collapse Across Scales

Tables 1 and 2 show that collapse occurs in both regimes. Despite model diversity in the shared synthetic pool, every family shows monotonic held-out perplexity growth across generations. The absolute rate varies by family: Llama degrades fastest, while Qwen, Gemma, and OLMo show a slower collapse.

Collapse also appears in lexical diversity. Table 3 reports distinct bigram and trigram rates on generated continuations. From generation 0 to generation 10, distinct-2 falls from 0.350 to 0.205 for Llama, 0.326 to 0.163 for OLMo, and 0.392 to 0.177 for Qwen. Distinct-3 shows the same pattern, falling by 35–55% depending on the family. The corresponding generated samples increasingly end in long list-like continuations; Appendix C gives representative examples.

Family	Distinct-2			Distinct-3		
	Gen0	Gen5	Gen10	Gen0	Gen5	Gen10
Llama	0.350	0.225	0.205	0.752	0.518	0.488
OLMo	0.326	0.194	0.163	0.727	0.445	0.357
Qwen	0.392	0.195	0.177	0.780	0.404	0.354

Table 3: Generated-text diversity during 7B–8B multi-model recursive training. Distinct-2 and distinct-3 are computed on generated continuations after stripping prompt tokens.

5.2 Task-Vector Negation Recovers PPL and Downstream Scores

We evaluate the one-step rewind in Eq. 4. For current generation i, this operation uses checkpoints i and $i+1$ to estimate a healthier checkpoint comparable to generation $i-1$. Table 4 shows that one-step rewind improves perplexity for every model family. In the full-parameter 1B–1.5B sweep, increasing α monotonically lowers PPL over the displayed range. The 8B QLoRA sweep also benefits from larger α for Llama and OLMo, while Qwen's optimum is near $\alpha = 1.5$.

The same intervention also improves downstream task metrics, though the effect is smaller than the perplexity recovery. Across the 8B downstream comparisons, the best recovered checkpoints close 54% of the average-score gap between the collapsed current checkpoint θ_i and the target checkpoint θ_{i-1}. Table 5 reports the corresponding average downstream scores. The recovered model moves toward the target but does not fully restore it, indicating that weight-space reversal is a partial repair rather than complete capability restoration.

Two-step extrapolation. If collapse were globally linear, the same local direction should also rewind multiple generations. We therefore apply Eq. 5 as a stricter diagnostic, comparing two estimators for the same earlier checkpoint θ_{i-2}: a checkpoint-matched one-step estimator that uses θ_{i-1} and θ_i, and the two-step estimator that uses only θ_i and θ_{i+1}. We select coefficients from collapse-step L2 norms alone (Appendix D gives the derivation and per-model values). Table 6 shows that two-step extrapolation is worse on perplexity on the held out C4 data and downstream tasks for both 1B-1.5B and 7B-8B multi model recovery. The local direction is informative, but extrapolating it farther backward is not a reliable mitigation.

Full-parameter 1B–1.5B multi-model recovery

Family	i	Target	Current	Future	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1.0$	$\alpha = 1.25$	$\alpha = 1.5$	$\alpha = 2.0$
Llama	2	25.55	51.10	102.96	45.65	41.49	38.35	35.90	33.96	32.46	**30.48**
Llama	4	102.96	172.59	250.94	159.51	149.17	139.40	130.52	123.31	116.80	**106.04**
Llama	6	250.94	335.81	431.18	317.50	300.05	284.66	270.73	257.46	245.15	**225.13**
Llama	8	431.18	535.87	682.15	505.60	479.08	454.03	429.51	408.78	388.51	**353.93**
Gemma	2	27.10	50.57	88.65	45.65	41.73	38.66	36.27	34.45	32.91	**30.83**
Gemma	4	88.65	134.76	182.04	126.06	118.10	110.89	104.18	97.96	92.92	**83.99**
Gemma	6	182.04	226.31	271.79	216.95	208.50	200.00	193.39	186.35	179.83	**168.18**
Gemma	8	271.79	316.48	365.45	307.04	295.80	285.85	276.90	268.12	259.00	**243.05**
Qwen	2	25.44	42.04	75.30	38.56	35.98	33.97	32.46	31.25	30.30	**29.01**
Qwen	4	75.30	118.26	166.72	110.22	103.33	97.46	92.39	88.02	84.14	**77.80**
Qwen	6	166.72	218.99	274.94	208.12	198.39	190.05	182.16	175.45	168.97	**157.82**
Qwen	8	274.94	346.74	411.40	332.98	320.35	309.19	298.49	288.34	279.43	**262.30**

QLoRA 7B–8B multi-model recovery

Family	i	Target	Current	Future	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1.0$	$\alpha = 1.25$	$\alpha = 1.5$	$\alpha = 2.0$
Llama	2	19.47	45.02	80.33	39.91	35.75	32.47	29.90	27.88	26.32	**24.10**
Llama	4	80.33	107.54	139.36	102.32	96.95	92.33	87.76	83.80	80.38	**74.24**
Llama	6	139.36	161.98	188.20	158.24	154.69	151.25	148.32	145.27	143.01	**138.78**
Llama	8	188.20	215.32	242.57	209.27	204.50	198.86	193.34	188.43	183.64	**174.30**
OLMo	2	21.82	32.74	44.63	30.77	29.11	27.68	26.49	25.48	24.66	**23.41**
OLMo	4	44.63	54.64	66.84	52.23	49.99	47.87	45.98	44.25	42.68	**39.93**
OLMo	6	66.84	78.79	89.41	76.33	73.82	71.45	69.19	66.90	64.80	**60.82**
OLMo	8	89.41	90.95	106.61	87.50	84.29	80.98	77.95	75.11	72.46	**67.49**
Qwen	2	20.13	27.90	34.77	26.98	26.27	25.79	25.41	25.18	**25.12**	25.43
Qwen	4	34.77	43.69	49.66	42.93	42.34	41.82	41.52	41.27	**41.20**	41.41
Qwen	6	49.66	59.49	66.86	58.23	57.35	56.56	55.80	55.44	55.08	**54.91**
Qwen	8	66.86	75.94	83.46	74.88	73.93	73.15	72.65	72.38	**72.27**	72.55

Table 4: Held-out C4 perplexity for one-step weight-space rewind across multi-model collapse trajectories. Columns: "Target" is the earlier checkpoint θ_{i-1} we aim to recover; "Current" is the collapsed checkpoint θ_i; "Future" is the one-step-further collapsed checkpoint θ_{i+1} used to form the negated update. Remaining columns sweep the rewind coefficient α in $\hat{\theta}_{i-1} = \theta_i - \alpha(\theta_{i+1} - \theta_i)$. Bold indicates the best α per row. Negation reduces perplexity for every family and checkpoint, with optimal α near 2.0 in most settings and near 1.5 for QLoRA Qwen3 8B.

5.3 Collapse Vectors Are Locally Structured

The recovery results suggest that collapse updates form a locally structured trajectory. To test this directly, we analyze consecutive collapse vectors

$$\tau_i = \theta_i - \theta_{i-1}. \tag{6}$$

If adjacent collapse updates were directionally unrelated, τ_{i+1} would be a poor proxy for moving from θ_i back toward θ_{i-1}. Instead, Table 7 shows that consecutive full-rank 1B–1.5B updates have consistently positive cosine similarity. The mean adjacent cosine is 0.210 for Llama, 0.204 for Gemma, and 0.303 for Qwen. When updates are compared two steps apart, the mean cosine falls to 0.149, 0.152, and 0.205, respectively. The 7B–8B effective-LoRA cosines are much smaller in absolute value, but remain positive on average. Figure 1 visualizes this task-vector alignment. Adjacent collapse task vectors have positive cosine similarity, so the next observed task vector points partly along the previous collapse direction and its negation is an informative local recovery direction.

Figure 2 shows the corresponding L2 geometry for both the full-rank 1B–1.5B and effective-LoRA 7B–8B settings. The full-rank collapse-step norms shrink smoothly across generations; for example, $\|\tau_i\|_2$ falls from 13.62 to 8.61 for Llama, 11.58 to 7.55 for Gemma, and 14.91 to 7.67 for Qwen. Because a later observed update is typically smaller than the earlier step we want to undo, this shrinkage motivates reverse coefficients above one. Figure 2 also plots the $\alpha = 1$ rewind residual $\|\tau_i - \tau_{i+1}\|_2$, which shows that two-step extrapolation always results in higher l2 norm compared to the one-step extrapolation. Combined with the imperfect cosine alignment above, this explains why one-step reversal can improve perplexity while two-step extrapolation remains unreliable (Appendix D): the update is informative locally, but its scale and direction are not stable enough to model the trajectory globally.

6 Discussion

Model diversity alone does not prevent collapse. Most empirical studies of language-model collapse use a single model that generates and retrains on its own outputs. One possible hypothesis is that generations from multiple model families would dilute

Full-parameter 1B–1.5B

Family	i	Target	Current	Future	$\alpha = 1.0$	$\alpha = 1.25$	$\alpha = 2.0$
Llama	3	57.30	56.97	56.27	56.90	**56.98**	56.67
Llama	5	56.27	55.57	54.72	**56.06**	56.02	56.01
Llama	7	54.72	53.98	52.58	54.59	54.69	**54.82**
Gemma	3	57.92	56.77	56.33	**57.03**	57.03	56.78
Gemma	5	56.33	56.23	55.90	**56.38**	56.26	56.04
Gemma	7	55.90	55.50	54.97	**55.51**	55.50	55.29
Qwen	3	60.54	59.84	59.43	59.89	**59.98**	59.93
Qwen	5	59.43	58.89	58.81	59.05	59.08	**59.09**
Qwen	7	58.81	58.37	57.93	**58.59**	58.47	58.39

QLoRA 7B–8B

Family	i	Target	Current	Future	$\alpha = 1.25$	$\alpha = 2.0$
Llama	2	72.07	70.97	69.66	**71.27**	70.99
Llama	4	69.66	68.80	68.32	**68.60**	68.42
Llama	8	67.71	67.44	67.32	**67.55**	67.31
OLMo	2	71.19	70.51	69.89	**70.96**	70.90
OLMo	4	69.89	69.47	68.98	69.78	**69.90**
OLMo	8	68.27	68.14	67.77	**68.40**	68.27
Qwen	2	72.14	71.36	70.59	**71.37**	70.46
Qwen	4	70.59	69.84	69.35	**69.88**	69.23
Qwen	8	68.74	68.45	68.05	**68.65**	68.06

Table 5: Average downstream `lm-evaluation-harness` accuracy for one-step weight-space rewind across multi-model collapse trajectories. Columns: "Target" is the earlier checkpoint θ_{i-1} we aim to recover; "Current" is the collapsed checkpoint θ_i; "Future" is the one-step-further collapsed checkpoint θ_{i+1} used to form the negated update. Remaining columns sweep the rewind coefficient α in $\hat{\theta}_{i-1} = \theta_i - \alpha(\theta_{i+1} - \theta_i)$. Scores are percentages averaged across the task suite (ARC-Challenge, ARC-Easy, BoolQ, HellaSwag, OpenBookQA, PIQA, SciQ, TruthfulQA-MC2, WinoGrande); higher is better. Bold marks the best recovered α per row. Recovered checkpoints close 54% of the average-score gap between the current and target checkpoints in the QLoRA 7B–8B setting, indicating partial but consistent capability restoration.

Setting	One-step PPL	Two-step PPL	Δ PPL	One-step DS (%)	Two-step DS (%)	Δ DS
Full-param 1B–1.5B	127.71	140.09	+12.37	57.75	57.29	-0.46
QLoRA 7B–8B	58.89	66.70	+7.81	66.70	66.66	-0.04

Table 6: Two-step extrapolation is worse than checkpoint-matched one-step recovery, indicating that the local collapse direction does not extend reliably across multiple generations. Both estimators target the same earlier checkpoint θ_{i-2}: the one-step estimator uses θ_{i-1} and θ_i, while the two-step estimator uses only θ_i and θ_{i+1} via Eq. 5. Coefficients are selected from collapse-step L2 norms alone (Appendix D). Means are over three source models and current checkpoints $i \in \{3, 7\}$, so both estimators target $\theta_{i-2} \in \{\theta_1, \theta_5\}$. We report perplexity (PPL) from held-out C4 dataset, downstream tasks (DS) from `lm-evaluation-harness`, and the delta between the one-step and two-step recovery methods.

Family	Adj. cos.	Dist.-2 cos.	$\lVert\tau_1\rVert_2$	$\lVert\tau_9\rVert_2$
Llama	0.210	0.149	13.62	8.61
Gemma	0.204	0.152	11.58	7.55
Qwen	0.303	0.205	14.91	7.67

Table 7: Geometry of full-rank 1B–1.5B multi-model collapse updates. The first cosine column reports mean adjacent similarity between τ_i and τ_{i+1}; the second reports mean distance-2 similarity between τ_i and τ_{i+2}. The lower distance-2 values support the locality of the collapse direction, while shrinking L2 norms indicate that later updates are smaller than earlier updates.

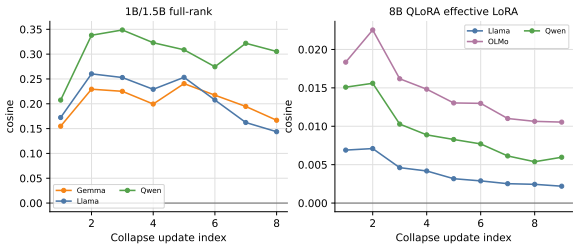

Figure 1: Cosine similarity between collapse task vectors. Adjacent task vectors are positively aligned, so the next observed collapse vector provides a useful local direction to negate.

biases of any single generator and slow or prevent collapse. Our results do not support this. Across both the 1B–1.5B full-rank and 7B–8B QLoRA settings, held-out perplexity grows monotonically and downstream accuracy degrades for every family, even though each model trains on a shuffled mixture of three families' outputs. This points toward different mitigation strategies—for example, mixing in real data (Gerstgrasser et al., 2024) or anchoring the loop to a fixed reference model that does not retrain on synthetic outputs (He et al., 2026)—rather than relying on generator diversity alone.

Collapse as directional drift. A useful way to read our recovery results is that the collapse trajectory in our setting behaves more like directional drift in weight space than like progressive destruction of capability. If collapsed checkpoints had simply lost the knowledge encoded in the pre-collapse model, no purely weight-space operation that uses only collapsed checkpoints could recover it. In-

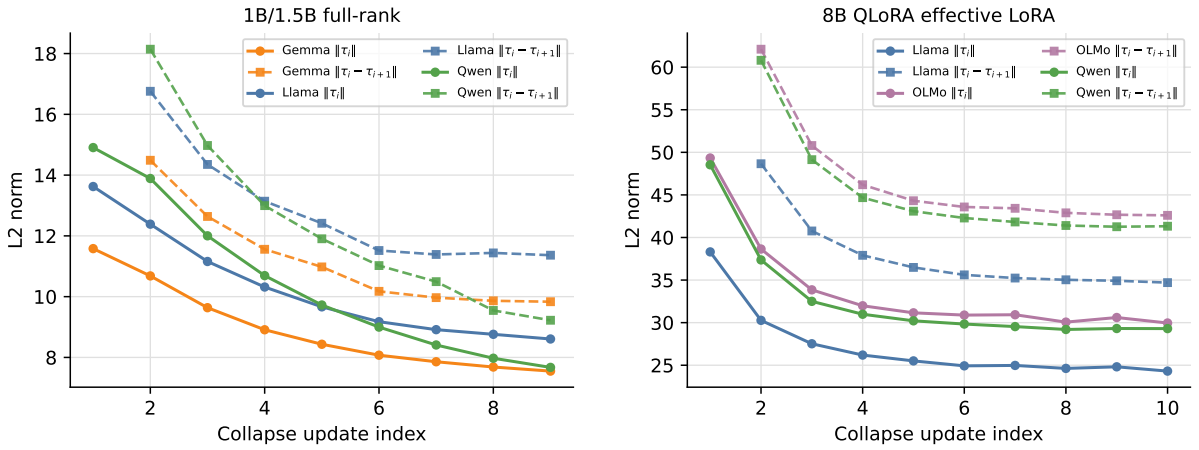

Figure 2: L2 geometry of the multi-model collapse trajectory motivates one-step rewind with $\alpha > 1$. Left: full-rank 1B–1.5B updates. Right: effective-LoRA 7B–8B updates, computed after converting each adapter to its dense LoRA update. Solid curves show collapse-step norms $\|\tau_i\|_2$, which shrink smoothly across generations: a later observed update is typically smaller than the earlier step we want to undo, so reverse coefficients above one are needed to match scale. Dashed curves show $\|\tau_i - \tau_{i+1}\|_2$, the L2 gap between the true previous update τ_i and the next observed update τ_{i+1} used as its proxy. The gap is nonzero and is relatively stable, confirming that negating the next collapse update is a local directional proxy.

stead, subtracting a scaled version of the most recent collapse update from the current checkpoint consistently reduces held-out perplexity and partially restores downstream accuracy, without ever consulting the pre-collapse model or any real data. The information needed to undo recent collapse is therefore present in the relationship between two collapsed checkpoints. We do not claim this picture extends to arbitrarily late collapse or to qualitatively different training regimes; the geometry analysis in §5.3 shows that adjacent collapse vectors are positively aligned but far from collinear, so the "drift direction" is only locally meaningful. Within that local regime, however, collapse appears to be a locally recoverable displacement.

Why perplexity recovers more than downstream accuracy. Held out C4 perplexity is recovered substantially by one step negation, especially in the later iterations, while downstream task accuracy is recovered only partially. We attribute this asymmetry primarily to alignment between the negated update and the loss landscape being measured. The negated update τ_{i+1} is the direction along which one more continued pretraining step on C4-like synthetic data moves the model, so it is most directly aligned with the C4 loss surface. Downstream Benchmarks such as ARC, HellaSwag, and BoolQ live on distributions where this direction is only partially aligned. A natural direction for future work is to test whether these results extend to recursive training setups in question-answering or instruction-following domains.

7 Limitations and Future Work

Scale and realism of the synthetic ecosystem. Our multi-model setup is a simplified proxy for the actual web. We use three model families per regime, a fixed generation recipe (temperature 0.7, top-p 0.9, repetition penalty 1.2), and a controlled shared pool with no human-written interleaving after the first generation. A more realistic web simulation would include more model families, varied decoding policies, ongoing injection of human-written text, and asymmetric model populations (e.g., a few large models dominating output volume). Additionally, the corpus we train on is small: roughly 100M tokens for initialization and 30M tokens per generation, several orders of magnitude below current web-scale pretraining (Grattafiori et al., 2024). The largest models we fine-tune are 8B-parameter models under QLoRA, model much smaller than the current state-of-the-art models. Full-parameter trajectories at the 8B+ scale would be a natural future direction.

Recovery requires checkpoint access and matched recipe. Although our method is data-free in the sense that no external real corpus is needed at recovery time, it is not assumption-free. The negation operation in Eq. 4 requires two adjacent checkpoints, θ_i and θ_{i+1}. Obtaining θ_{i+1}

in practice requires running one additional round of synthetic-data fine-tuning with a recipe similar to the one that produced θ_i. In settings where the original training recipe is not disclosed, the recovered direction may not align well with the collapse direction we wish to undo.

Simplicity of the merging operation. We use the minimal form of task-vector negation. More sophisticated merging operations, including TIES-merging (Yadav et al., 2023), tangent-space task arithmetic (Ortiz-Jimenez et al., 2023), and consensus-based negation (Kim et al., 2025), are likely complementary and may close more of the downstream gap, especially in the QLoRA setting where parameter-efficient adapter geometry differs from full-parameter updates (Zhang et al., 2023). We did not optimize for the strongest possible reversal operator; our goal was to test whether the local collapse direction itself is informative. That a simple linear negation already produces consistent recovery suggests that more refined operators are a promising direction.

Pretraining-style fine-tuning, not instruction-tuned use. Our experiments use base models with continued pretraining rather than the instruction-tuned and RLHF-trained variants that dominate public LLM usage. The full pretrain $\rightarrow$ SFT $\rightarrow$ preference-optimization pipeline would be more representative of deployed systems, but executing it iteratively across multiple generations and multiple model families is beyond scope available to us. Whether the collapse trajectory and its local reversibility behave the same way under instruction tuning, or under contamination of preference data rather than pretraining data, remains open.

Single-seed main results. The main multi-model collapse and recovery results (Tables 1, 2, 4, 5) are reported from a single seed per setting due to compute constraints; the single-model trajectory in Appendix A includes four to five seeds and shows that the collapse-rate variance is moderate. Replicating the multi-model runs with seed variance is the most direct robustness check we did not perform due to compute. We plan to complete additional multi-seed runs of the multi-model collapse scenario by the review and discussion phase.

Evaluation breadth. Held-out C4 perplexity is sensitive to distributional fit on a single corpus, and the lm-evaluation-harness suite measures narrow knowledge and reasoning capabilities. Downstream evaluations of generation quality, factuality, calibration, and long-form coherence would give a fuller picture of what is and is not recovered by weight-space reversal. We see consistent but small downstream improvements in our setting, but we cannot claim that the recovered models are functionally equivalent to the earlier target checkpoints.

8 Conclusion

We studied model collapse in a shared multi-model synthetic data loop, a setting closer to the likely future web than single-model self-training: many model families generate text, their outputs mix into a common corpus, and future models train on that mixture. Across full-rank 1B–1.5B models and QLoRA-trained 7B–8B models, multi-model pooling does not prevent collapse—perplexity grows monotonically and downstream accuracy degrades for every family. We then showed that recent collapse updates contain a recoverable local direction. Negating the next synthetic-training update reduces held-out perplexity for every family and partially recovers downstream accuracy, without using any external data. Trajectory analysis explains both effects through a single geometric picture: adjacent collapse vectors are positively but imperfectly aligned, and their norms shrink across generations. The shrinking norms mean that the next observed update is typically smaller than the earlier step we want to undo, which is why reversal coefficients $\alpha > 1$ recover best; the imperfect alignment is why two-step extrapolation, which compounds both the scale mismatch and the directional drift, degrades considerably. Taken together, these results suggest that within a local regime, multi-model collapse behaves like a directional displacement in weight space, one that can be partially undone using only checkpoints from the collapse loop itself.

References

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

Sina Alemohammad, Ahmed Imtiaz Humayun, Shruti Agarwal, John Collomosse, and Richard Baraniuk. 2024. Self-improving diffusion models with synthetic data. *Preprint*, arXiv:2408.16333.

Sina Alemohammad, Zhangyang Wang, and Richard Baraniuk. 2026. Neon: Negative extrapolation from

self-training improves image generation. In *The Fourteenth International Conference on Learning Representations*.

Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. 2025. Strong model collapse. In *The Thirteenth International Conference on Learning Representations*.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. 2024. A tale of tails: model collapse as a change of scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Jonas Dolezal, Sawood Alam, Mark Graham, and Maty Bohacek. 2026. The impact of ai-generated text on the internet. *Preprint*, arXiv:2604.26965.

Junhao Dong, Hao Zhu, Yifei Zhang, Xinghua Qu, Yew Soon Ong, and Piotr Koniusz. 2025. Machine unlearning via task simplex arithmetic. In *Advances in Neural Information Processing Systems*, volume 38, pages 166676–166707. Curran Associates, Inc.

George Drayson, Emine Yilmaz, and Vasileios Lampos. 2025. Machine-generated text detection prevents language model collapse. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29657–29673, Suzhou, China. Association for Computational Linguistics.

Yunzhen Feng, Elvis Dohmatob, Pu Yang, Francois Charton, and Julia Kempe. 2025. Beyond model collapse: Scaling up with synthesized data requires verification. In *The Thirteenth International Conference on Learning Representations*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

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. 2024. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *Preprint*, arXiv:2404.01413.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. 2026. Self-distillation zero: Self-revision turns binary rewards into dense supervision. *Preprint*, arXiv:2604.12002.

Zizhao Hu, Mohammad Rostami, and Jesse Thomason. 2025. Multimodal synthetic data finetuning and model collapse: Insights from vlms and diffusion models. In *Proceedings of the 27th International Conference on Multimodal Interaction*, ICMI '25, page 588–599, New York, NY, USA. Association for Computing Machinery.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Joshua Kazdan, Rylan Schaeffer, Apratim Dey, Matthias Gerstgrasser, Rafael Rafailov, David L. Donoho, and Sanmi Koyejo. 2025. Collapse or thrive? perils and promises of synthetic data in a self-generating world.

Hyoseo Kim, Dongyoon Han, and Junsuk Choe. 2025. Negmerge: Consensual weight negation for strong machine unlearning.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2026. Olmo 3. *Preprint*, arXiv:2512.13961.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. 2023. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Tianyu Wang, Akira Horiguchi, Lingyou Pang, and Carey E. Priebe. 2025. Llm web dynamics: Tracing model collapse in a network of llms. *Preprint*, arXiv:2506.15690.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Bingji Yi, Qiyuan Liu, Yuwei Cheng, and Haifeng Xu. 2026. Escaping model collapse via synthetic data verification: Near-term improvements and long-term convergence. In *The Fourteenth International Conference on Learning Representations*.

Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. 2023. Composing parameter-efficient modules with arithmetic operations. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Xuekai Zhu, Daixuan Cheng, Hengli Li, Kaiyan Zhang, Ermo Hua, Xingtai Lv, Ning Ding, Zhouhan Lin, Zilong Zheng, and Bowen Zhou. 2025. How to synthesize text data without model collapse? In *Forty-second International Conference on Machine Learning*.

Target gen.	Source gens.	Target PPL	Following PPL	Recovered PPL
2	θ_3, θ_4	65.07	183.45	95.04
3	θ_4, θ_5	183.45	505.21	256.68
4	θ_5, θ_6	505.21	1246.98	668.41
5	θ_6, θ_7	1246.98	2711.18	1531.01

Table 8: Single-model full-rank perplexity recovery for Qwen 2.5 1.5B. A recovered target generation g is computed from source checkpoints θ_{g+1} and θ_{g+2} as $\hat{\theta}_g = \theta_{g+1} - (\theta_{g+2} - \theta_{g+1})$. Recovered uses $\alpha = 1.0$.

Target gen.	Checkpoint	Source gens.	HellaSwag	ARC-Easy	ARC-Chal.	PIQA	Avg.
0	Target θ_0	–	66.91	72.18	43.26	75.41	64.44
2	Target θ_2	–	64.82	69.61	42.49	75.41	63.08
2	Following θ_3	–	61.76	69.57	41.55	74.76	61.91
2	Recovered $\hat{\theta}_2$	θ_3, θ_4	66.01	70.33	43.09	75.30	63.68
4	Target θ_4	–	59.29	69.28	41.13	74.76	61.11
4	Following θ_5	–	57.61	69.40	41.89	74.48	60.85
4	Recovered $\hat{\theta}_4$	θ_5, θ_6	61.13	68.98	41.30	74.86	61.57
6	Target θ_6	–	56.46	69.40	42.06	74.21	60.54
6	Following θ_7	–	55.82	69.40	42.41	74.16	60.45
6	Recovered $\hat{\theta}_6$	θ_7, θ_8	57.37	69.15	42.06	74.43	60.75

Table 9: Single-model Qwen 2.5 1.5B downstream collapse and recovery. Values are normalized accuracies reported as percentages. A recovered target generation g is computed as $\hat{\theta}_g = \theta_{g+1} - (\theta_{g+2} - \theta_{g+1})$ and compared against the target checkpoint θ_g.

A Single-Model Collapse and Reversal

The main paper focuses on multi-model synthetic feedback, but single-model recursive training is a useful supporting regime for collapse and recovery diagnostics. Figure 3 shows the full Qwen 2.5 1.5B full-rank trajectory across five seeds and the Qwen 2.5 3B QLoRA bridge experiment across four seeds. Because the single-model and multi-model full-rank runs used different effective data scales, these curves should not be interpreted as a matched collapse-rate comparison. The 3B results are included only as a bridge collapse regime between the 1.5B full-rank and 7B–8B QLoRA settings. In the full-rank single-model Qwen 2.5 1.5B loop, perplexity rises from 21.07 at generation 0 to 5375.83 at generation 7. Table 8 shows that the same one-step negation idea also reduces perplexity in this single-model setting. Table 9 shows the same comparison for downstream tasks. Each recovery block names the target generation directly: the recovered checkpoint is produced from the following generation and its one-step-further collapsed checkpoint, then compared against that target. For example, the recovered generation 2 row is produced from generations 3 and 4.

B Multi-Model Reversal Graphs

The main text reports reversal results primarily in tables so that the target, current, future, and recovered checkpoints can be compared directly. For completeness, Figures 5 and 6 show the corresponding multi-model reversal plots.

C Generated Samples Across Collapse

Table 10 shows middle and late excerpts from Llama 8B generated texts across generations. Because each saved example contains the prompt followed by the generated continuation, we report middle and late excerpts rather than only the beginning of the stored text. The examples are intended to illustrate the qualitative failure mode behind the distinct-ngram drop in Table 3. We see that the collapse is gradual. Early generations often remain locally document-like in the middle, while later generations increasingly drift into lists, malformed lists, or repeated boilerplate.

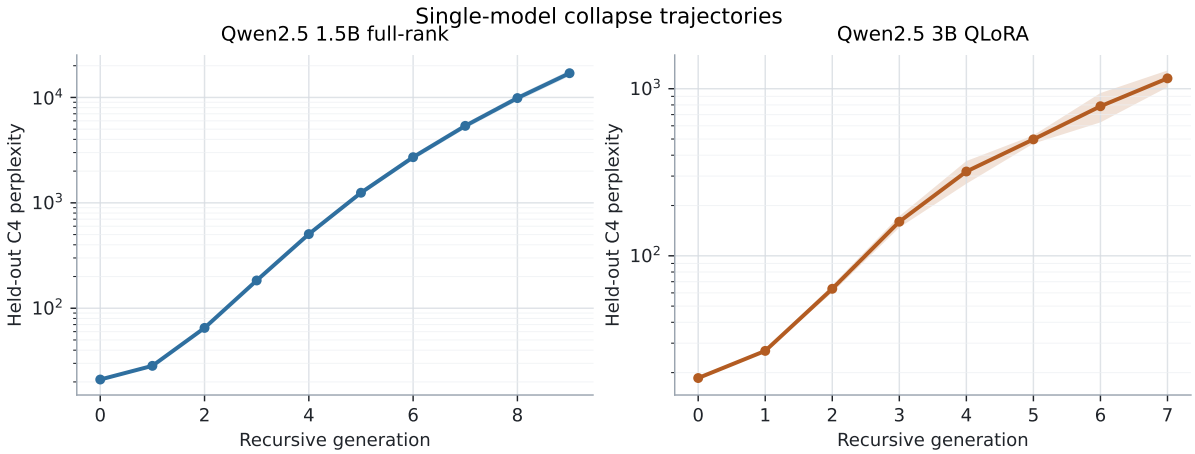

Figure 3: Single-model collapse trajectories with cross-seed standard deviation. Left: Qwen 2.5 1.5B full-rank self-training across five seeds. Right: Qwen 2.5 3B QLoRA self-training across four seeds, using 150k training examples per generation. The 3B experiment serves as a bridge between the 1.5B full-rank and 8B QLoRA regimes.

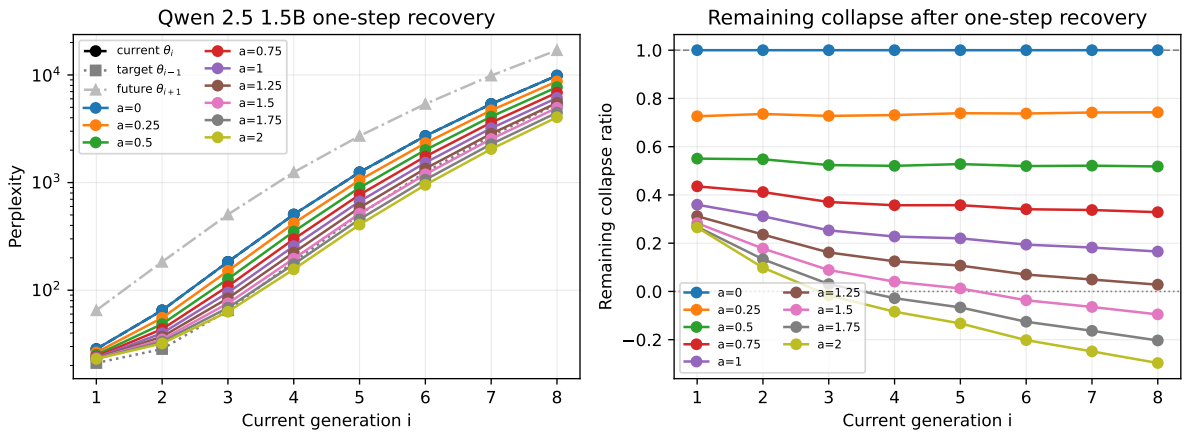

Figure 4: Single-model Qwen 2.5 1.5B full-rank collapse reversal. Each comparison uses the current checkpoint θ_i and the next collapsed checkpoint θ_{i+1} to estimate the previous checkpoint θ_{i-1} by task-vector negation.

D Multi-Step Rewind Is Noisy

The main experiments use one-step rewind. We also evaluate the two-step closed-form extrapolation in Eq. 5, which attempts to move from θ_i toward θ_{i-2} using only the most recent observed direction $\theta_{i+1} - \theta_i$. This test asks whether a single local collapse direction remains reliable when extrapolated farther backward. With coefficients selected only from collapse-step L2 norms, we see that two-step rewind is consistently worse on perplexity and approximately tied, but slightly worse, on the downstream subset.

Derivation. First consider the generic current checkpoint θ_i. The one-step estimator assumes that the next observed collapse update can approximate the previous collapse update:

$$\hat{\theta}_{i-1} = \theta_i - \alpha(\theta_{i+1} - \theta_i). \tag{7}$$

To extrapolate one additional step backward, we apply the same local-update assumption recursively. Starting from the one-step estimate, the hypothetical earlier checkpoint is

$$\begin{aligned} \hat{\theta}_{i-2} &= \hat{\theta}_{i-1} - \alpha(\theta_i - \hat{\theta}_{i-1}) && (8)\\ &= [\theta_i - \alpha(\theta_{i+1} - \theta_i)] - \alpha\left(\theta_i - [\theta_i - \alpha(\theta_{i+1} - \theta_i)]\right) && (9)\\ &= \theta_i - (\alpha + \alpha^2)(\theta_{i+1} - \theta_i). && (10) \end{aligned}$$

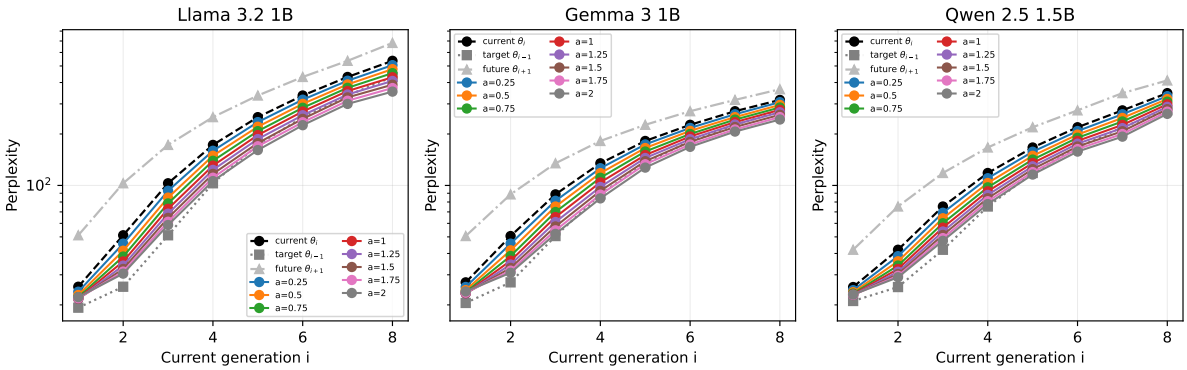

Figure 5: Full-parameter 1B–1.5B multi-model collapse reversal. Recovered checkpoints are compared with the earlier target checkpoint and the collapsed current checkpoint for each source model.

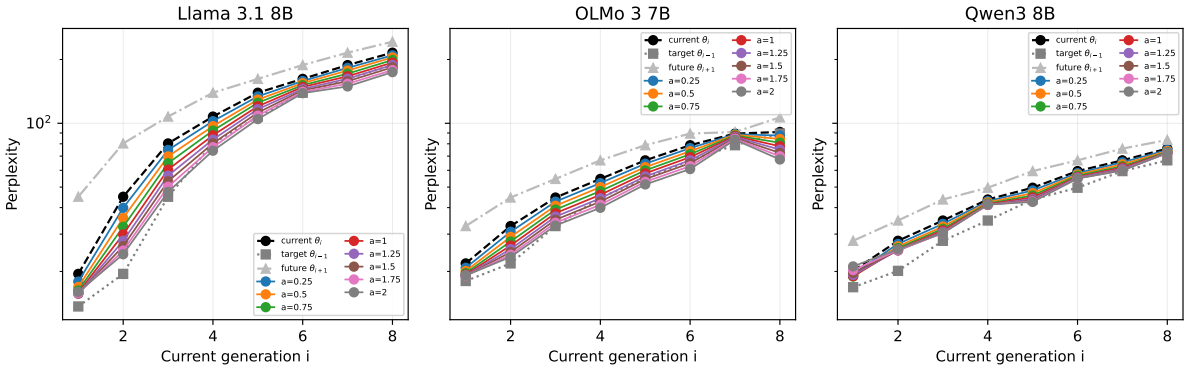

Figure 6: QLoRA 7B–8B multi-model collapse reversal. The same one-step task-vector negation procedure improves held-out perplexity relative to the collapsed current checkpoint in the larger-model setting.

Thus the two-step diagnostic is equivalent to applying an effective coefficient $\beta = \alpha + \alpha^2$ to the single observed future update. In this derivation, θ_i is the current checkpoint for the two-step estimator and the intended target is θ_{i-2}. For the checkpoint-matched diagnostic below, one-step recovery is shifted back by one checkpoint so that both estimators target the same parameter, θ_{i-2}. Thus when we report two-step current checkpoints $i = 3$ and $i = 7$, both estimators target θ_1 and θ_5, respectively.

Table 6 in the paper summarizes this comparison; Figures 7 and 8 break it down by source model and current checkpoint. For each source model, we test two-step current checkpoints $i \in \{3, 7\}$. The matched one-step estimator uses θ_{i-1} and θ_i to estimate θ_{i-2}, while the two-step estimator uses θ_i and θ_{i+1} to estimate the same θ_{i-2}. Concretely, for the $i = 3$ comparison, one-step uses θ_2 and θ_3 to estimate θ_1, while two-step uses θ_3 and θ_4 to estimate the same θ_1. For the $i = 7$ comparison, one-step uses θ_6 and θ_7 to estimate θ_5, while two-step uses θ_7 and θ_8.

We choose the comparison coefficients from L2 norms alone. Let $\tau_j = \theta_j - \theta_{j-1}$ be the collapse update from generation $j-1$ to generation j. For the checkpoint-matched one-step estimator, which estimates θ_{i-2} from θ_{i-1} and θ_i, we use

$$\alpha_{m,i}^{\text{1step}} = \frac{\|\tau_{i-1}^{(m)}\|_2}{\|\tau_i^{(m)}\|_2}. \tag{11}$$

For the two-step estimator, the target displacement from θ_i to θ_{i-2} has L2 scale approximated by $\|\tau_{i-1}^{(m)}\|_2 + \|\tau_i^{(m)}\|_2$, while the observed future update is $\tau_{i+1}^{(m)}$. We therefore set the effective two-step multiplier to

$$\beta_{m,i}^{\text{2step}} = \frac{\|\tau_{i-1}^{(m)}\|_2 + \|\tau_i^{(m)}\|_2}{\|\tau_{i+1}^{(m)}\|_2}. \tag{12}$$

Gen	Middle continuation excerpt	Late continuation excerpt
0	included on this CD, try searching software archive sites. See Also: Hard Drive Maintenance Discs …	He holds an MBA from Durham University Business School and is married with three children …
1	This position requires you work full-time hours Monday–Friday from early morning until late evening …	We offer competitive salaries commensurate with experience level plus a generous benefits package …
3	We reserve right not accept any guest who does NOT meet minimum age requirement …	rest assured peace of mind knowing full confidence satisfaction guaranteed end result achieved …
5	they don't have any service plan active under those names anymore …	contained within document body text sections included below table formatting style layout design elements …
7	works closely alongside producer/engineer/arranger/songwriter/musical director/film composer …	online resources offered schools colleges universities worldwide making learning fun exciting engaging relevant practical …
9	all rooms while still maintaining privacy when needed …	Rebrand Image Brand Identity Marketing Communications Campaign Promotion Sales Funnel Optimization SEO SEM PPC …

Table 10: Representative middle and late excerpts from Llama 8B multi-model generations. Early generations remain locally coherent for longer, while later generations drift into list-like or generic continuations.

Setting	Model	i	Target	$\alpha^{1\text{step}}$	$\beta^{2\text{step}}$
Full-param 1B–1.5B	Llama	3	θ_1	1.098	2.259
Full-param 1B–1.5B	Llama	7	θ_5	1.029	2.068
Full-param 1B–1.5B	Gemma	3	θ_1	1.103	2.271
Full-param 1B–1.5B	Gemma	7	θ_5	1.030	2.076
Full-param 1B–1.5B	Qwen	3	θ_1	1.159	2.407
Full-param 1B–1.5B	Qwen	7	θ_5	1.069	2.175
QLoRA 7B–8B	Llama	3	θ_1	1.100	2.206
QLoRA 7B–8B	Llama	7	θ_5	0.998	2.026
QLoRA 7B–8B	OLMo	3	θ_1	1.142	2.267
QLoRA 7B–8B	OLMo	7	θ_5	0.999	2.056
QLoRA 7B–8B	Qwen	3	θ_1	1.150	2.255
QLoRA 7B–8B	Qwen	7	θ_5	1.010	2.033

Table 11: L2-only coefficients for the checkpoint-matched one-step and two-step comparison. $\alpha^{1\text{step}}$ scales the observed one-step collapse update used to estimate θ_{i-2} from θ_{i-1} and θ_i. $\beta^{2\text{step}}$ is the effective multiplier applied to $\theta_{i+1} - \theta_i$ in Eq. 5.

We apply this value directly as the effective multiplier in Eq. 5. This is an oracle best-case diagnostic: although $\|\tau_{i-1}^{(m)}\|_2$ may be unavailable in the intended recovery setup, where we begin from $\theta_i^{(m)}$, we assume access to it here to give two-step extrapolation a favorable scale choice. Table 11 lists these per-model coefficients. For downstream tasks, we use `lm-evaluation-harness` with `-limit 0.1`, so the downstream averages are computed on a 10% subset of each task. With these L2-only coefficients, two-step extrapolation is consistently worse than one-step recovery on perplexity and roughly tied but slightly worse on downstream accuracy. The collapse direction is locally useful, but extrapolating it across multiple generations is not a reliable mitigation.

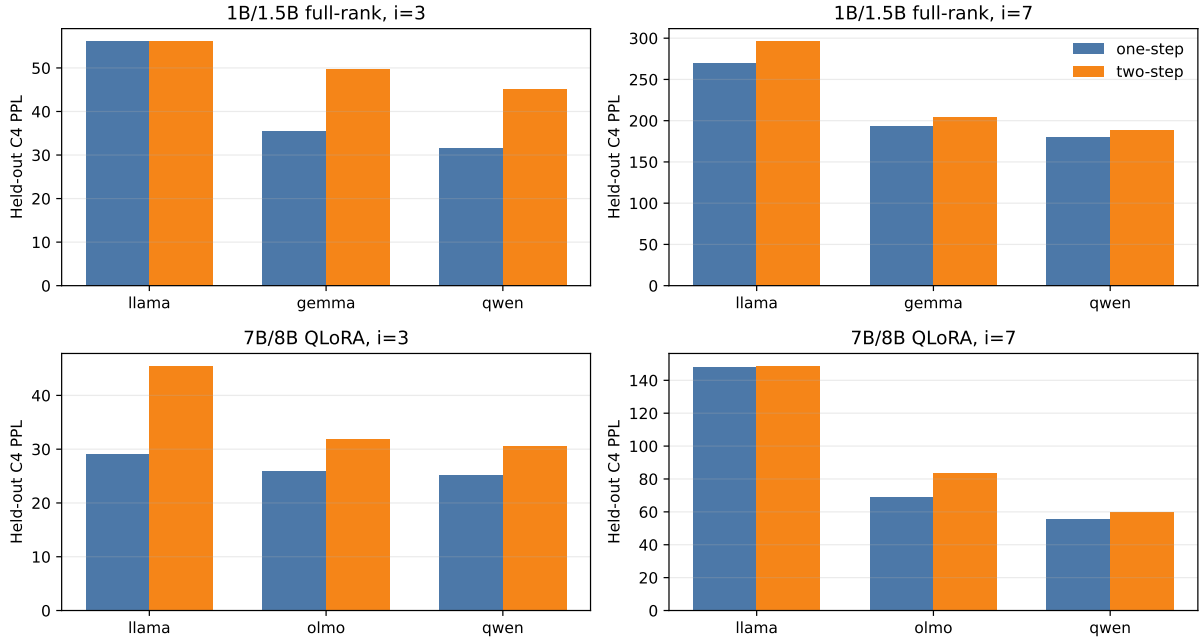

Figure 7: Checkpoint-matched perplexity comparison for L2-selected one-step and two-step recovery, separated by setting, source model, and current checkpoint $i \in \{3, 7\}$. In this diagnostic, $i = 3$ targets θ_1 and $i = 7$ targets θ_5.

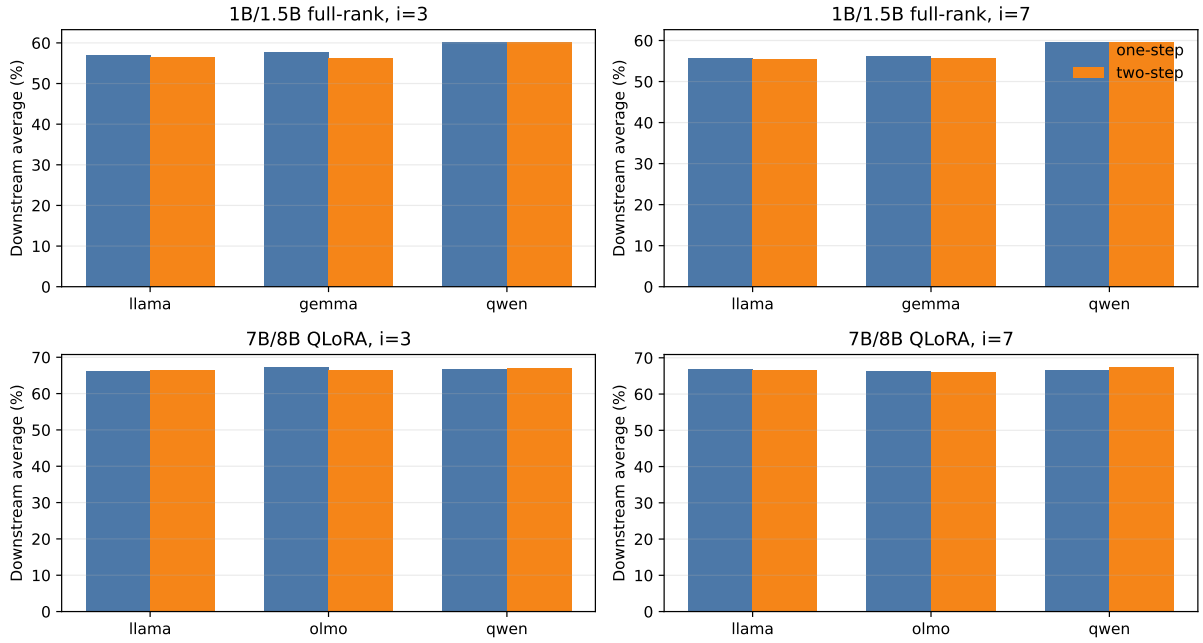

Figure 8: Checkpoint-matched downstream comparison for L2-selected one-step and two-step recovery, separated by setting, source model, and current checkpoint $i \in \{3, 7\}$. Scores are averaged over a 10% `lm-evaluation-harness` subset.