RANK-1 FISHER FOR CONTINUAL LEARNING: WHEN IS ITS GEOMETRY FAITHFUL

Anonymous authors
Paper under double-blind review

ABSTRACT

Mean-gradient rank-1 empirical-Fisher surrogates offer a memory-efficient approach to continual learning, but their practical utility does not establish that one retained direction faithfully represents loss-gradient second moments on unseen data. We examine this geometric interpretation theoretically and empirically. The motivating linear-model proof omits an input outer product: the gradient is proportional to $(A - \gamma I)xx^\top$, rather than $\|x\|^2(A - \gamma I)$. Under the corresponding linear–Gaussian model, we derive the exact population second-moment operator and show that, although $\lambda_2/\lambda_1 = 2/(d+2)$ vanishes with dimension for a scaled-identity residual, the optimal rank-1 relative Frobenius error approaches $\sqrt{2/3}$. Simulations further show that same-sample evaluation can spuriously favor rank-1 at low n/d, whereas this advantage disappears on disjoint test samples. Held-out evaluations on three frozen masked diffusion language models spanning 219M to 8B parameters reveal no uniform rank-1 advantage: at the largest calibration size, rank-1 slightly outperforms diagonal on SMDM-219M, but underperforms on SMDM-1.14B and in 14 of 15 LLaDA-8B settings. An exact error decomposition gives a diagnostic answer: held-out faithfulness jointly requires rank-1 capacity, transfer of the mean-gradient direction, and transfer of its fitted scale. A source-setting check on small MNIST UNets provides a favorable local boundary, but the overall results vary across objectives, scales, layers, and corruption levels. Mean-gradient rank-1 Fisher therefore need not provide a faithful held-out reconstruction, although this does not rule out its consolidation utility. Code is available.

1 INTRODUCTION

Continual learning studies how models can acquire new tasks without catastrophically overwriting earlier knowledge (McCloskey & Cohen, 1989; French, 1999; Parisi et al., 2019). Two recurring responses are to revisit prior behavior through replay and to constrain parameter updates using importance estimates from earlier tasks (Shin et al., 2017; Zenke et al., 2017; Aljundi et al., 2018). Both approaches become harder to scale as models and task sequences grow: replay must preserve useful past behavior, while parameter regularization must compress which changes are safe.

Elastic weight consolidation (EWC) makes the second problem geometric. It uses a Fisher-weighted quadratic penalty to discourage movement along directions that matter to previous tasks (Kirkpatrick et al., 2017). The matrix approximation therefore determines which parameter displacements are protected, rather than merely storing a list of importance scores. Full parameter-space second moments are intractable for modern models; diagonal EWC discards cross-parameter correlations, while structured or rotated approximations retain more geometry at additional cost (Ritter et al., 2018; Liu et al., 2018). A rank-1 surrogate $cuu^\top$ occupies the opposite extreme: one cheap cross-parameter direction is retained, and every orthogonal displacement is left unpenalized.

This compression question has recently reached continual diffusion models, where generative replay is already a natural tool (Gao & Liu, 2023; Masip et al., 2025). An ICLR 2026 study combines replay with rank-1 EWC and argues that low signal-to-noise ratio (SNR) makes per-example gradients collinear with their mean (Wang et al., 2026). Under that argument, a mean-gradient outer product acts as an effectively rank-1 empirical Fisher at a cost comparable to diagonal EWC. The method reports lower forgetting than replay-only and diagonal-EWC baselines in continual image generation,

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.

making the proposed geometry a practically consequential explanation rather than only a matrix approximation.

The application result and the geometric explanation nevertheless require different evidence. The conventional "empirical Fisher" is an observed loss-gradient second moment and need not equal the model Fisher or Hessian (Kunstner et al., 2019). A rank-1 penalty can help because its chosen direction interacts favorably with replay or task updates even when it reconstructs that second moment poorly; conversely, a small matrix-approximation error cannot by itself guarantee retention. Establishing rank-1 geometry therefore requires a reconstruction criterion that accounts for the full residual spectrum, not only an eigengap (Eckart & Young, 1936).

We narrow this gap through one theoretical and one statistical question. First, do the stated linear–Gaussian assumptions actually imply rank-1 gradient second-moment geometry? Second, does a mean-gradient surrogate fitted from a finite gradient sample reconstruct an independent sample from the same population? Reusing the fitted sample cannot answer the second question: the empirical Fisher and mean direction then share the same examples, a dependence that is especially consequential when the slice dimension d greatly exceeds the number of gradients n.

Finally, our primary empirical study tests whether any surviving structure transfers from continuous image diffusion to masked diffusion language models (MDLMs). This is a boundary test: Gaussian corruption and denoising MSE are replaced by categorical token masking and importance-weighted cross-entropy (Austin et al., 2021; Lou et al., 2024; Sahoo et al., 2024; Nie et al., 2025a). Mask probability is therefore an operational corruption coordinate rather than Gaussian SNR. Appendix Section B.1 separately applies our held-out protocol to the source paper's small MNIST UNet; it checks geometry in that setting without re-evaluating its continual image-generation results. Throughout, "empirical Fisher" refers only to the observed loss-gradient second moment.

Contributions.

- **Theoretical boundary.** Retaining the $xx^\top$ factor missing from the source derivation yields a generally input-dependent gradient. We derive the resulting population second-moment operator and give a closed-form counterexample in which $\lambda_2/\lambda_1 \to 0$ while the optimal rank-1 Frobenius residual remains ≈ 0.816.
- **Estimation boundary.** An isotropic null reproduces strong same-sample rank-1 preference at low n/d even though the population second moment is the identity. A disjoint test sample removes it in every simulated case.
- **Transfer boundary.** A source-setting MNIST check finds a rank-1 advantage in 26/27 held-out seed–timestep cells, whereas across SMDM-219M, SMDM-1.14B, and LLaDA-8B the comparison depends on checkpoint, measured layer, mask probability, and calibration split. Target-count, dense-slice, and cross-text controls further delimit this transfer; an exact decomposition separates scalar miscalibration, direction mismatch, and the residual that no rank-1 matrix can remove. Nested test-size and common-update audits respectively show that finite test-Fisher noise does not explain the larger-model ordering and that Frobenius and update-direction fidelity can rank approximations differently.

Together these results delimit the Fisher-geometric interpretation of a continual-learning-motivated surrogate. They do not re-evaluate the reported continual image-generation outcomes or claim that Frobenius error predicts update utility; they identify which theoretical implication fails, which finite-sample comparison is misleading, and where rank-1 has lower held-out error under the measured masked-language conditions.

2 Related Work

Continual learning and Fisher approximations. EWC uses a Fisher-weighted quadratic penalty to discourage movement in parameters important to previous tasks (Kirkpatrick et al., 2017). SI instead accumulates loss-changing parameter trajectories, while MAS measures output sensitivity (Zenke et al., 2017; Aljundi et al., 2018). Natural-gradient and Kronecker-factored methods motivate retaining structured Fisher curvature (Amari, 1998; Martens & Grosse, 2015); in continual learning, structured online Laplace keeps blockwise Kronecker factors, while rotated EWC seeks coordinates in which a diagonal Fisher is more informative (Ritter et al., 2018; Liu et al., 2018). These methods

establish that the chosen curvature structure can matter for continual learning. Most directly, the reference method combines generative replay with rank-1 EWC and reports lower forgetting than replay-only and diagonal variants (Wang et al., 2026). We take those results as motivation for testing the surrogate's geometric premise, not as evidence that utility and fidelity are equivalent.

Low-rank Fisher evaluation. The outer product of observed loss gradients is usually called the empirical Fisher, although it need not equal the model Fisher or the Hessian away from special conditions (Kunstner et al., 2019). We preserve that conventional name but study only the gradient second moment. The best rank-k matrix approximation under Frobenius norm is characterized by the Eckart–Young theorem (Eckart & Young, 1936); randomized decompositions are a practical route when the matrix is too large to materialize (Halko et al., 2011). Our held-out top-eigen rank-1 fit is therefore a capacity lower bound, not a deployable estimator.

Masked diffusion language models. Continuous diffusion commonly uses Gaussian corruption and denoising objectives (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021). Discrete and masked language diffusion instead corrupts token categories, including D3PM, SEDD, MDLM, SMDM, and LLaDA (Austin et al., 2021; Lou et al., 2024; Sahoo et al., 2024; Nie et al., 2025a;b). Their categorical cross-entropy objectives do not inherit continuous linear–Gaussian gradient geometry by construction, making them a transfer test for the rank-1 interpretation rather than a replication setting.

3 Theoretical Analysis

3.1 Gradient Correction

Consider the linear denoiser used in the source argument (Wang et al., 2026). At a fixed diffusion time, let $s_A(x) = Ax$, let the target be γx, and absorb the time-dependent loss weight into $\kappa > 0$:

$$\ell(A; x) = \frac{\kappa}{2}\|(A - \gamma I)x\|_2^2. \tag{1}$$

Writing $B = A - \gamma I$, differentiation with respect to the entries of A gives

$$g_A(x) = \nabla_A \ell(A; x) = \kappa B x x^\top. \tag{2}$$

The displayed source proof instead replaces this matrix by $\kappa\|x\|_2^2 B$. For the scaled-identity residual $B = \beta I$, $\beta \neq 0$, that equality would require $xx^\top = \|x\|_2^2 I$, which is false for nonzero x when $d > 1$. Consequently, even this linear scaled-identity case does not make finite-residual per-example gradients scalar multiples of one fixed matrix.

This correction does not assert that an empirical diffusion model cannot exhibit low-rank gradients. It shows that the stated linearity and low-SNR assumptions do not establish that conclusion.

3.2 Spectrum and Rank-1 Error

Let $x \sim \mathcal{N}(0, \sigma^2 I_d)$ and identify parameter-space vectors with $d \times d$ matrices under the Frobenius inner product. Define the population loss-gradient second-moment operator $\mathcal{F}_B$ by

$$\mathcal{F}_B[H] = \mathbb{E}[\langle g_A(x), H\rangle_F g_A(x)]\,. \tag{3}$$

Proposition 1. *For the loss in Eq. (1),*

$$\mathcal{F}_B[H] = \kappa^2\sigma^4 B\left[\mathrm{tr}(B^\top H)I + B^\top H + H^\top B\right]. \tag{4}$$

Proof. Eq. (2) gives $\langle g_A(x), H\rangle_F = \kappa x^\top B^\top H x$. Isserlis' identity for an isotropic Gaussian yields (Isserlis, 1918)

$$\mathbb{E}\left[(x^\top C x)xx^\top\right] = \sigma^4\left[\mathrm{tr}(C)I + C + C^\top\right]. \tag{5}$$

Substitute $C = B^\top H$ and left-multiply by $\kappa^2 B$. *QED*

The operator is not generally rank-1. A particularly revealing case is the scaled-identity residual assumed by the motivating intuition.

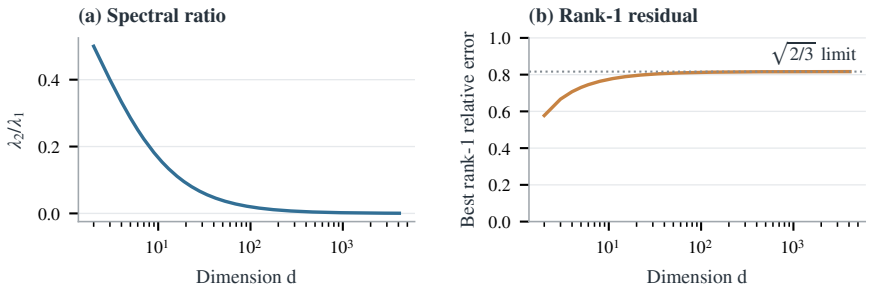


Figure 1: Closed-form counterexample within the motivating linear–Gaussian setting. For $B = \beta I$, $\beta \neq 0$, the second-to-first eigenvalue ratio vanishes (a), while the best rank-1 relative Frobenius residual approaches $\sqrt{2/3}$ (b). The growing multiplicity of the $2C$ eigenspace prevents rank-1 reconstruction. Curves are exact analytic values, so no sampling interval is shown.

Corollary 1. *Let $B = \beta I_d$ with $\beta \neq 0$, and set $C = \kappa^2\beta^2\sigma^4$. The eigenvalues of $\mathcal{F}_B$ are*

$$C(d+2) \quad \textit{on } \operatorname{span}\{I\}, \qquad \textit{multiplicity } 1, \tag{6}$$

$$2C \quad \textit{on traceless symmetric matrices}, \qquad \textit{multiplicity } \frac{d(d+1)}{2} - 1, \tag{7}$$

$$0 \quad \textit{on antisymmetric matrices}, \qquad \textit{multiplicity } \frac{d(d-1)}{2}. \tag{8}$$

Thus

$$\frac{\lambda_2}{\lambda_1} = \frac{2}{d+2}, \qquad \frac{\|\mathcal{F}_B - (\mathcal{F}_B)_1\|_F}{\|\mathcal{F}_B\|_F} = \sqrt{\frac{2(d-1)}{3d}}. \tag{9}$$

Moreover, $\mathbb{E}[g_A(x)] = \kappa\beta\sigma^2 I$, exactly the leading eigendirection in this example. Thus an optimally scaled mean-gradient outer product attains the best possible rank-1 approximation and still leaves the residual in Eq. (9). The ratio λ_2/λ_1 tends to zero, but the multiplicity of the second eigenvalue grows quadratically. The optimal rank-1 residual therefore tends to $\sqrt{2/3} \approx 0.816$ (Fig. 1). An eigengap and a small Frobenius residual are different properties. At exact convergence ($B = 0$), all gradients and the Fisher in this idealized loss are zero; there is then no nontrivial shared direction to retain.

4 Evaluation Framework

4.1 Split-Sample Fisher Reconstruction

For a parameter slice $\theta_S \in \mathbb{R}^d$, let $g_i = \nabla_{\theta_S}\ell_i$ be a per-example gradient. Calibration and test sets define independent empirical Fishers

$$\widehat{F}_{\text{cal}} = \frac{1}{n_c}\sum_{i=1}^{n_c} g_i g_i^\top, \qquad \widehat{F}_{\text{test}} = \frac{1}{n_t}\sum_{j=1}^{n_t} h_j h_j^\top. \tag{10}$$

All estimators below are functions of calibration gradients only.

The mean-gradient estimator uses $\mu = n_c^{-1}\sum_i g_i$ and fits the least-squares coefficient along its outer product:

$$R_\mu = c_\mu \mu\mu^\top, \qquad c_\mu = \frac{\mu^\top \widehat{F}_{\text{cal}}\mu}{\|\mu\|_2^4}. \tag{11}$$

The matched diagonal estimator is $D = \operatorname{diag}(\widehat{F}_{\text{cal}})$. The leading eigenpair of *the test Fisher* is reported only as an oracle lower bound on rank-1 error, not as an available estimator.

For any fitted surrogate Q, the held-out endpoint is

$$e_{\text{test}}(Q) = \frac{\|\widehat{F}_{\text{test}} - Q\|_F}{\|\widehat{F}_{\text{test}}\|_F}. \tag{12}$$

This is the primary experimental endpoint; lower values indicate a closer reconstruction of the test-gradient second moment. The reported metrics and their aggregation are specified in Section 5.5.

4.2 Direction, Scale, and Rank-1 Capacity

Let $u = \mu/\|\mu\|_2$ and write the fitted mean-gradient surrogate as $R_\mu = a_{\text{cal}} u u^\top$, where $a_{\text{cal}} = u^\top \widehat{F}_{\text{cal}} u$. Holding its direction fixed, the test-optimal coefficient is

$$a^*_{\text{test}} = u^\top \widehat{F}_{\text{test}} u, \qquad R^*_\mu = a^*_{\text{test}} u u^\top. \tag{13}$$

Define $e_{\text{fit}} = e_{\text{test}}(R_\mu)$, $e_{\text{rescale}} = e_{\text{test}}(R^*_\mu)$, and let e_{oracle} be the test-Fisher oracle rank-1 error. Optimality along u and over all rank-1 matrices gives the exact squared-error decomposition

$$e^2_{\text{fit}} = \underbrace{e^2_{\text{fit}} - e^2_{\text{rescale}}}_{\text{scale}} + \underbrace{e^2_{\text{rescale}} - e^2_{\text{oracle}}}_{\text{direction}} + \underbrace{e^2_{\text{oracle}}}_{\text{rank-1 capacity}}. \tag{14}$$

We also report $|u^\top v_1|$, where v_1 is a leading eigenvector of $\widehat{F}_{\text{test}}$. The residual terms in Eq. (14) are the primary diagnosis; alignment is a complementary direction diagnostic.

Update-weighted fidelity. Frobenius error weights every direction, whereas an actual step probes only its realized displacement. For a common update $\delta\theta$, we report

$$r_\delta(Q) = \frac{|\delta\theta^\top (\widehat{F}_{\text{test}} - Q)\delta\theta|}{\delta\theta^\top \widehat{F}_{\text{test}} \delta\theta + \epsilon}, \quad q_\delta(Q) = \frac{\delta\theta^\top Q \delta\theta}{\delta\theta^\top \widehat{F}_{\text{test}} \delta\theta + \epsilon}, \tag{15}$$

with $\epsilon = 10^{-30}$. These are the relative directional-energy error and estimated-to-test energy ratio; neither identifies a causal effect on forgetting.

4.3 Finite-Sample Null

The direction μ is built from the same rows that form $\widehat{F}_{\text{cal}}$. When $n_c \ll d$, this algebraic dependence can make R_μ look competitive on $\widehat{F}_{\text{cal}}$ even if it has no population advantage. We quantify the effect under the deliberately unfavorable null $g_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$, whose population second moment is exactly diagonal and isotropic.

We simulate 200 repetitions at $d = 768$ with $n_c \in \{8, 16, 32, 64, 128\}$ and at $d = 1792$ with $n_c \in \{8, 16, 32, 64\}$. Independent test sizes are 128 and 64, respectively, matching our MDLM slices. At low n_c/d, same-sample scoring selects rank-1 in all 200 repetitions for most cases. Independent-test and known-population scoring favor diagonal in all 200 repetitions for every case (Fig. 2). This null does not model language gradients; it isolates a protocol-level disagreement between the two evaluation targets.

5 Experimental Setup

5.1 Masked-Language Objective

The supplied SMDM training code samples $t \sim U[0, 1]$, sets $p = (1 - \varepsilon)t + \varepsilon$ with $\varepsilon = 10^{-3}$, independently masks tokens, and uses importance-weighted cross-entropy (Nie et al., 2025a). For a length-L token sequence x and mask m, our per-example loss is

$$\ell(x, m, p) = \frac{1}{pL} \sum_{j=1}^{L} m_j \, \text{CE}(f_\theta(x \odot (1 - m), m)_j, x_j). \tag{16}$$

We evaluate five fixed probabilities $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Empty masks are resampled; at $p = 0.1$ and $L = 64$ the excluded event has probability $0.9^{64} \approx 0.0012$, and it is smaller at larger p. Mask probability is a corruption coordinate, not Gaussian SNR.

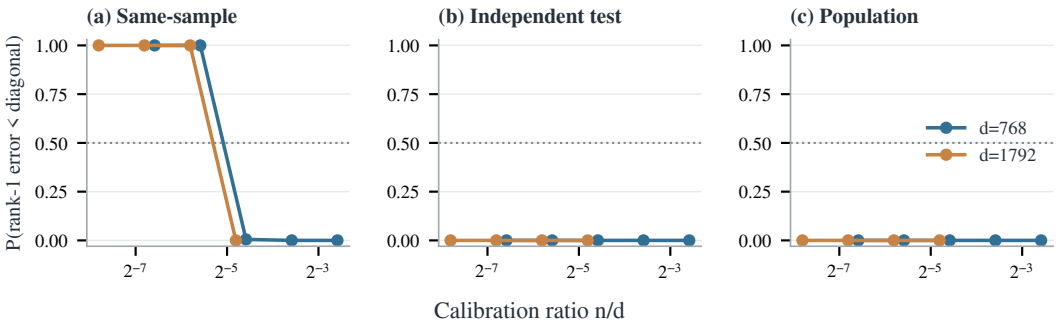


Figure 2: Finite-sample rank-1 preference under an isotropic null. Curves show the probability, over 200 repetitions, that the fitted mean-gradient rank-1 error is below the fitted diagonal error. Reusing the calibration Fisher for scoring creates near-certain wins at low n/d (left). A disjoint test Fisher (middle) and the known population target (right) favor diagonal in every simulated case.

Table 1: Primary split-sample design. Calibration prefixes reuse one fixed gradient pool within a seed; every prefix is scored on the same disjoint test pool.

model	**parameters**	**slice** d	**layers**	**seeds**	$n_c \mid n_t$
SMDM-219M	219,050,496	768	0, 8, 17	5	32/64/128 | 128
SMDM-1.14B	1,142,367,744	1792	0, 9, 19	3	16/32/64 | 64
LLaDA-8B	8,015,581,184	4096	0, 15, 31	5	16/32/64 | 64

5.2 Models, Parameters, and Data Splits

We use frozen SMDM checkpoints with 219,050,496 and 1,142,367,744 parameters and the public LLaDA-8B-Base checkpoint with 8,015,581,184 parameters (Nie et al., 2025a;b). We probe early, middle, and late attention normalization scale vectors (Table 1). On SMDM-219M, matched layer-0 and layer-17 attention-output matrices provide dense-slice controls.

The text source is the official GSM8K test split, using question and answer text but not task labels or generated answers (Cobbe et al., 2021). Each document is tokenized independently with the checkpoint tokenizer and truncated to 64 tokens. For each seed, source documents are shuffled once; the first $\max n_c$ form nested calibration prefixes and the next n_t form the test set. Examples and masks are therefore disjoint between fitting and scoring.

5.3 Target-Count Intervention

Increasing p changes at least two things: the visible context and the number of supervised targets. To separate them, we replay exactly the same corrupted contexts but retain one uniformly selected corrupted target per sequence:

$$\ell_{\text{one}}(x, m, p) = \frac{1}{pL}\text{CE}(f_\theta(x \odot (1-m), m)_J, x_J), \qquad J \sim \text{Uniform}\{j : m_j = 1\}. \tag{17}$$

Probe seeds 0–2 use identical calibration and test examples, masks, and $64 \mid 64$ sample sizes in the all-target and one-target runs. A validation record checks source-line/token hashes and mask hashes for each pair. At fixed p, any common scalar rescaling would cancel from relative Frobenius geometry; the intervention changes which token gradients are summed.

5.4 Test-Size and Common-Update Audits

To test whether a finite test Fisher determines the ranking, we freeze one seed-0, $n_c = 64$ calibration surrogate per model and score it on nested $n_t \in \{16, 32, 64, 128, 256\}$ prefixes over the 15-cell grid,

using five test seeds for SMDM-219M and three for each larger model. At $n_t = 256$ the source pool is shared and masks vary; this is not a refit of the 219M $n_c = 128$ endpoint.

Separately, both surrogates score the same full-parameter update from three unregularized GD/replay adaptations of the SMDM-1.14B GSM8K checkpoint on two Dolly tasks (1,000 updates each). Its held-out Fisher uses 64 masked examples from 32 source problems excluded from calibration and replay; neither surrogate produces the update.

5.5 Metrics and Aggregation

Our primary metric is the held-out relative Frobenius error $e_{\text{test}}(Q)$ in Eq. (12). We report it directly for the mean-gradient rank-1 surrogate R_μ, the diagonal surrogate D, and the test-Fisher rank-1 oracle. The oracle uses test gradients and is only a capacity lower bound; it is not an available estimator. For the same-sample diagnostic in Fig. 3, we analogously replace $\widehat{F}_{\text{test}}$ by $\widehat{F}_{\text{cal}}$ in both the numerator and denominator.

Cell maps report the paired difference $\Delta = e_{\text{test}}(D) - e_{\text{test}}(R_\mu)$, so positive values favor rank-1. Relative errors can exceed one when a fitted surrogate is farther from $\widehat{F}_{\text{test}}$ than the zero matrix. Lower error than diagonal establishes only a relative advantage, not a faithful reconstruction.

For each model and calibration size, we first average direct errors over the fixed 15-cell layer–mask grid within each split/mask seed, then report the mean and sample standard deviation of those seed means. We also count the number of cell means for which $\Delta > 0$. These descriptive summaries are conditional on one checkpoint per model; no hypothesis test is applied to the fixed grid, and we report no asymptotic p-values.

For the direction–scale diagnostic, fitted, rescaled, oracle, and alignment values use the same seed-grid aggregation. The three shares in Eq. (14) are normalized within each paired cell and then averaged over the stated fixed grid and seeds.

For the test-size audit, direct errors are first averaged over the 15-cell grid within each test seed, then summarized across seeds. The common-update audit reports Eq. (15) across the three seeds.

6 Results

6.1 Effect of Sample Reuse and Test Size

Fig. 3 compares calibration and test scoring on real GSM8K gradients. For SMDM-219M, calibration favors rank-1 at every n_c, whereas held-out scoring favors diagonal at $n_c = 32$ before favoring rank-1 at 64 and 128. For SMDM-1.14B, calibration again favors rank-1 throughout, but held-out scoring favors diagonal throughout. LLaDA-8B favors diagonal under both scoring rules, while held-out errors reveal much greater split/mask-seed sensitivity. Thus independent scoring can reverse a surrogate ranking or expose instability hidden by calibration error; increasing n_c reduces held-out error without making rankings uniform across models.

The nested-prefix audit fixes $n_c = 64$. Aggregate ordering stabilizes by $n_t = 32$ for SMDM-219M and by $n_t = 16$ for both larger models. At $n_t = 256$, rank-1/diagonal errors are $.869/.877$, $12.13/4.93$, and $.955/.724$ for SMDM-219M, SMDM-1.14B, and LLaDA-8B, respectively. The 219M advantage changes sign in one of five test seeds and wins only 5/15 cell means. Hence finite test-Fisher noise does not explain the larger-model ordering (Appendix Table 5).

6.2 Variation across Models, Layers, and Masks

At maximum n_c, grid-averaged error is lower for rank-1 on SMDM-219M even though rank-1 wins only 6 of 15 layer–mask cells; both the grid average and cell counts favor diagonal on SMDM-1.14B and LLaDA-8B (Table 2). The LLaDA estimate is highly split/mask-seed sensitive: one layer-0, $p = 0.3$ cell in seed 2 has rank-1 error 801.94 even at $n_c = 64$; it is retained in every summary. The test-sample oracle is much lower for all models, showing that the fitted mean-gradient surrogate does not attain the best rank-1 approximation of those test samples. This gap alone does not identify

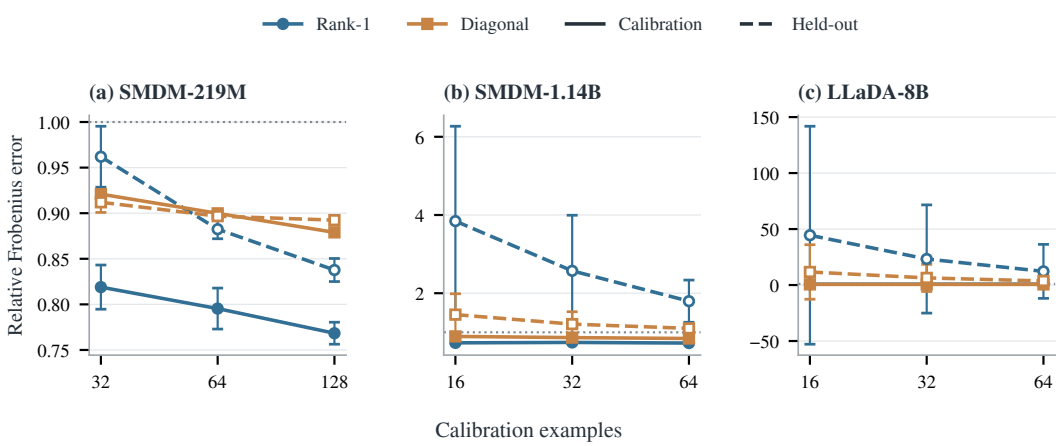


Figure 3: Same fitted surrogates, different scoring samples. Solid lines score fitted surrogates on their calibration Fisher; dashed lines score on disjoint examples and masks. Curves are mean direct relative Frobenius errors; error bars are one sample SD across split/mask seed means.

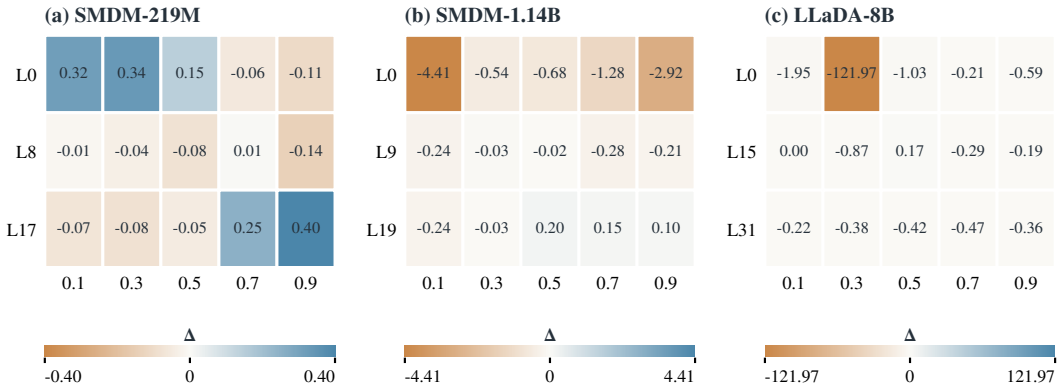


Figure 4: **Paired direct-error difference $e_{\text{diag}} - e_{\text{rank1}}$ at maximum n_c.** Each cell averages split/mask seeds; positive values (blue) favor rank-1. Panels use separate color scales because the LLaDA estimates include an extreme value; values are printed in every cell. Horizontal axis: mask probability; vertical axis: layer index.

whether direction, scale, or rank-1 capacity is responsible; the decomposition below separates them on the audit-qualified controls that retain the required sufficient statistics.

6.3 Failure Anatomy and Scope Controls

Direction, scale, and capacity. At $n_c = n_t = 64$, the matched SMDM-219M controls comprise 45 seed–layer–mask cells for each target-count objective and nine cells for the late dense slice at layer 17. Table 3 reports the fitted error, the error after replacing only its scalar coefficient by the test-optimal value, the test-oracle rank-1 error, and alignment with the test leading direction.

Test-optimal rescaling improves all 45 all-target cells, all 45 one-target cells, and all nine late-dense cells, but it does not close the oracle gap. Averaging the normalized squared-error components in Eq. (14), the scale/direction/capacity shares are 12/51/37% for all targets, 11/51/38% for one target, and 4/22/74% for the late dense slice. Thus scalar miscalibration is present but is not the main limitation: normalization-slice error is dominated by direction mismatch, whereas the measured dense slice is dominated by the residual that no rank-1 matrix can remove. Supervising one target further weakens direction alignment.

Table 2: Direct held-out relative Frobenius error at maximum n_c, reported as mean±sample SD over split/mask seed grid means. "Oracle" is the best rank-1 fit to the test Fisher and is not an available estimator.

model	n_c	rank-1	diagonal	oracle	rank-1 cells
SMDM-219M	128	0.838 ± 0.013	0.892 ± 0.003	0.531 ± 0.022	6/15
SMDM-1.14B	64	1.797 ± 0.539	1.101 ± 0.036	0.449 ± 0.046	3/15
LLaDA-8B	64	12.212 ± 24.098	3.627 ± 5.991	0.270 ± 0.032	1/15

Table 3: Direction–scale diagnostics at $n_c = n_t = 64$, reported as mean±sample SD over seed grid means.

setting	fitted	rescaled	oracle	$\lvert u^\top v_1 \rvert$
All targets	$.956 \pm .083$	$.860 \pm .018$	$.540 \pm .015$	$.571 \pm .059$
One target	$1.202 \pm .192$	$.970 \pm .007$	$.599 \pm .016$	$.286 \pm .022$
Late dense	$1.005 \pm .003$	$.982 \pm .009$	$.860 \pm .014$	$.265 \pm .112$

A common realized update. On the same full-parameter GD/replay update, rank-1 has lower r_δ in all three seeds ($.761 \pm .244$ versus $.997 \pm .003$). Both underestimate held-out Fisher energy: q_δ is $.239 \pm .244$ and $.0029 \pm .0025$, respectively. Because this audit differs from the slice study in checkpoint and scope, it shows metric/scope dependence rather than a controlled reversal or a fidelity–forgetting correlation (Appendix Table 6).

Scope controls. The same protocol favors rank-1 in 26/27 cells on small MNIST UNets (Appendix Section B.1), showing that it can recover favorable source-setting structure. On SMDM attention-output matrices, diagonal wins five of six 219M cells and rank-1 wins two of six 1.14B cell means; both remain near error one (Appendix Fig. 8 and Table 7).

Holding examples, masks, and contexts fixed, one-target supervision changes Δ from -0.029 ± 0.066 to -0.215 ± 0.154; the paired shift is positive in 12/15 cells, but three seeds do not support a model-wide effect (Appendix Fig. 7).

Appendix Section B reports cross-text and end-to-end controls; full-test GSM8K retention shows no stable Rank-1+GD advantage over trace-matched Diagonal+GD and does not make reconstruction error a forgetting predictor.

7 Limitations

Results use one checkpoint per model and mostly selected slices; the test-size audit shares its $n_t = 256$ records, and the update audit has 64 held-out examples and three observational seeds. It identifies neither a fidelity–forgetting relationship nor continual image-generation utility.

8 Conclusion

Across theory, an isotropic null, and held-out MDLM gradients, rank-1 reconstruction is not universal: sample reuse can create a spurious advantage, and larger test Fishers preserve model-dependent ordering. Direction mismatch dominates audited normalization slices and rank-1 capacity the measured dense slice; a common-update audit can nevertheless favor rank-1, while the MNIST and GSM8K controls show only local geometry and no stable rank-1 retention advantage. Rank-1 geometry is faithful when rank-1 capacity, direction transfer, and scale transfer all hold on unseen gradients; our experiments find this combination only locally, not as a general consequence of low SNR, model scale, or continual-learning utility.

AI USE STATEMENT

Generative AI tools assisted with developing and checking mathematical arguments and code, providing feedback on experimental design and interpretation, searching and formatting references, and drafting and editing the manuscript. Language-model-generated replay targets were used only in the continual-learning controls described in the paper. The authors reviewed all AI-assisted text and code, checked citations against primary sources, and verified reported claims against derivations and raw experiment artifacts. The authors take responsibility for the final content.

REPRODUCIBILITY STATEMENT

The assumptions and proofs appear in Section 3 and Appendix A; the estimators, data splits, controls, and aggregation rules are specified in Sections 4 to 6. Appendix C records software versions, commands, artifact counts, hashes, audit checks, and invalidated-run handling. Checked manifests and summarizers recompute the reported values from experiment artifacts; Appendix C identifies separately validated audit families.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *European conference on computer vision*, pp. 144–161. Springer, 2018.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Anonymous Authors. Stabilizing language models under continual learning via condition-anchored distillation, 2027. Concurrent submission to the International Conference on Learning Representations.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2107.03006.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". *arXiv preprint arXiv:2309.12288*, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Rei Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Databricks. Databricks dolly 15k. https://huggingface.co/datasets/databricks/databricks-dolly-15k, 2023. Dataset card and data repository.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3 (4):128–135, 1999.

Rui Gao and Weiwei Liu. DDGR: Continual learning with deep diffusion-based generative replay. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 10744–10763. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23e.html.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011. doi: 10.1137/090771806.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

L. Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1–2):134–139, 1918. doi: 10.1093/biomet/12.1-2.134.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems*, 2019.

Xialei Liu, Marc Masana, Luis Herranz, Joost Van de Weijer, Antonio M Lopez, and Andrew D Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. *arXiv preprint arXiv:1802.02950*, 2018.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 32819–32848. PMLR, 2024. URL https://proceedings.mlr.press/v235/lou24a.html.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 2408–2417. PMLR, 2015. URL https://proceedings.mlr.press/v37/martens15.html.

Sergi Masip, Pau Rodriguez, Tinne Tuytelaars, and Gido M. van de Ven. Continual learning of diffusion models with generative distillation. In *Proceedings of the 3rd Conference on Lifelong Learning Agents*, volume 274, pp. 431–456. PMLR, 2025. URL https://proceedings.mlr.press/v274/masip25a.html.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Academic Press, 1989.

Shen Nie, Fengqi Zhu, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. Scaling up masked diffusion models on text. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=WNvvwK0tut.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025b.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured laplace approximations for overcoming catastrophic forgetting. *Advances in Neural Information Processing Systems*, 31, 2018.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.07524.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, 2017.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Zekun Wang, Anant Gupta, Zihan Dong, and Christopher J. MacLellan. Avoid catastrophic forgetting with rank-1 fisher from diffusion models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ae90d88755e0eaeb9121712fbac4e8de-Abstract-Conference.html.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. Pmlr, 2017.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

A Matrix-Free Evaluation

Let $G_c \in \mathbb{R}^{n_c \times d}$ and $G_t \in \mathbb{R}^{n_t \times d}$ contain calibration and test gradients as rows. Their Fishers are $G_c^\top G_c/n_c$ and $G_t^\top G_t/n_t$. The test Frobenius norm is available from its Gram matrix:

$$\|\widehat{F}_{\text{test}}\|_F^2 = \operatorname{tr}\left[(G_t G_t^\top/n_t)^2\right]. \tag{18}$$

For $R = cuu^\top$ fitted on calibration data,

$$\langle \widehat{F}_{\text{test}}, R\rangle_F = \frac{c}{n_t}\|G_t u\|_2^2, \tag{19}$$

$$\|R\|_F^2 = c^2\|u\|_2^4. \tag{20}$$

For unit u, the test-optimal coefficient used in Eq. (14) is $a^*_{\text{test}} = \|G_t u\|_2^2/n_t$. A leading test-Fisher direction is recovered from a leading eigenvector of $G_t G_t^\top/n_t$, so both $|u^\top v_1|$ and the three decomposition terms are likewise available from Gram and cross-Gram products without materializing a $d \times d$ matrix. The square on c in Eq. (20) is essential. An earlier internal run omitted it; all artifacts affected by that bug are explicitly invalidated and excluded from this paper.

For the diagonal estimator, let $d_c = n_c^{-1}\sum_i g_i \odot g_i$ and define d_t analogously. Then

$$\|\widehat{F}_{\text{test}} - \operatorname{diag}(d_c)\|_F^2 = \|\widehat{F}_{\text{test}}\|_F^2 - 2d_t^\top d_c + \|d_c\|_2^2. \tag{21}$$

Nonzero eigenvalues of either empirical Fisher equal those of its example Gram matrix, so the held-out oracle also requires no $d \times d$ materialization. For a fixed update $\delta\theta$, the three directional energies used in Eq. (15) are likewise matrix-free:

$$\delta\theta^\top \widehat{F}_{\text{test}}\delta\theta = \frac{1}{n_t}\sum_j (h_j^\top \delta\theta)^2, \quad \delta\theta^\top R\delta\theta = c(u^\top\delta\theta)^2, \quad \delta\theta^\top D\delta\theta = d_c^\top(\delta\theta \odot \delta\theta). \tag{22}$$

The probe self-check additionally constructs small Fishers explicitly and verifies the mean-rank-1 and diagonal formulas and the ordering $e_{\text{oracle rank1}} \le e_{R_\mu}$. The submission manifest also checks $e_{\text{oracle rank1}} \le e_{R_{\text{top}}}$ in every submitted row.

B Additional Results and Controls

B.1 Source-Setting Check in Continuous Image Diffusion

We separately apply the split-sample reconstruction test in the source paper's small MNIST diffusion setting (Wang et al., 2026). Using its released implementation, we train three independently initialized 152,497-parameter UNets for 200 epochs on the full MNIST training split. At each fixed diffusion timestep $t \in \{100, 200, \ldots, 900\}$, we fit the mean-gradient rank-1 and diagonal surrogates to 1,024 test-split examples and score them on 1,024 disjoint examples with independent diffusion noise. The oracle is the best rank-1 fit to each scoring Fisher. All measurements use the full parameter gradient rather than a selected slice.

Rank-1 has lower held-out error than diagonal in 26 of 27 seed–timestep cells; the exception is seed 0 at $t = 500$. Averaged equally over this grid, the held-out errors are 0.644 for rank-1, 0.994 for diagonal, and 0.179 for the oracle. The calibration and held-out rank-1 averages are similar (0.637 and 0.644), so this advantage is not removed by independent scoring at $n_c = n_t = 1{,}024$. The timestep trend is also consistent with stronger low-SNR rank-1 structure: mean held-out rank-1 error falls from 0.899 ± 0.143 at $t = 100$ to 0.214 ± 0.325 at $t = 900$, while diagonal remains near 0.994. However, the large cross-seed variation and the persistent gap to the oracle show that the mean-gradient direction is not a uniformly faithful rank-1 estimator. This check concerns Fisher reconstruction in the small source model; it does not re-evaluate continual image-generation performance.

B.2 Test-Size and Common-Update Audits

Table 5 reports every nested test prefix used in the fixed-$n_c = 64$ audit. The values are descriptive over these tested prefixes: "stable" in the main text means that the aggregate winner does not change at any subsequent listed prefix, not that a population convergence theorem has been established. At

Table 4: MNIST UNet held-out relative Frobenius error. Values are mean±sample SD over three independent training seeds; the final column counts seeds with lower rank-1 than diagonal error.

t	**rank-1**	**diagonal**	**oracle**	**rank-1 seeds**
100	0.8995 ± 0.1427	0.9951 ± 0.0019	0.5116 ± 0.1393	3/3
200	0.9458 ± 0.0753	0.9944 ± 0.0037	0.3173 ± 0.0321	3/3
300	0.8956 ± 0.0666	0.9938 ± 0.0029	0.2335 ± 0.0520	3/3
400	0.8870 ± 0.1251	0.9934 ± 0.0010	0.1622 ± 0.0413	3/3
500	0.7321 ± 0.3271	0.9934 ± 0.0006	0.0881 ± 0.0061	2/3
600	0.5386 ± 0.4114	0.9938 ± 0.0007	0.0561 ± 0.0239	3/3
700	0.4377 ± 0.3775	0.9941 ± 0.0006	0.0608 ± 0.0583	3/3
800	0.2480 ± 0.3222	0.9943 ± 0.0005	0.0819 ± 0.1124	3/3
900	0.2142 ± 0.3248	0.9944 ± 0.0006	0.0994 ± 0.1502	3/3

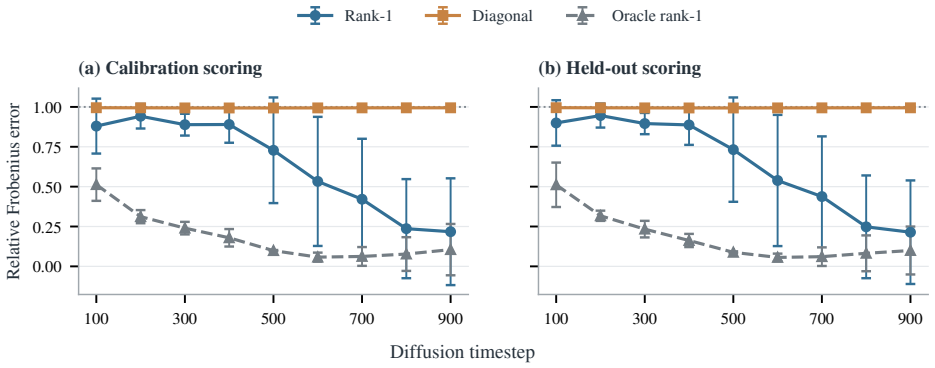


Figure 5: **Source-setting MNIST UNet check.** Relative Frobenius error under calibration scoring (left) and disjoint held-out scoring (right). Points and error bars are mean±sample SD over three independent training seeds. The oracle uses the scoring Fisher and is not an available estimator.

$n_t = 256$, all replicates use the same exhausted source record pool with independently sampled masks.

The common-update audit evaluates both surrogates on the same unregularized full-parameter GD/replay displacement. Rank-1 is closer in all three seeds, but both estimates lie below the held-out Fisher energy in every seed (Table 6). The matched held-out source loss also increases in every run (0.311 ± 0.207), but three observational pairs cannot support a fidelity–forgetting correlation.

B.3 Dense-Slice Control at 1.14B

We extend the dense-slice control to SMDM-1.14B using the same held-out estimand. The run family was fixed before seeing its GPU results: layers 0 and 19, mask probabilities $0.1/0.5/0.9$, three seeds, and $n_c = n_t = 64$ disjoint examples. Each attention-output matrix has $d = 3{,}211{,}264$ parameters. All 18 seed–layer–mask cells are retained.

Across the six-cell grid, the rank-1 and diagonal errors are 1.224 ± 0.305 and 1.00008 ± 0.00027 (sample SD over three seed grid means); the oracle error is 0.610 ± 0.016. The paired seed-level score $\log(e_{\text{diag}}/e_{\text{rank1}})$ is -0.106 ± 0.070 SEM, corresponding to a geometric-mean $e_{\text{rank1}}/e_{\text{diag}}$ ratio of 1.112. Rank-1 wins 8/18 seed-cells but only 2/6 cell means. We retain the seed-0, layer-0, $p = 0.1$ rank-1 error of 4.456, which visibly affects its cell mean. Thus the dense control now spans both SMDM checkpoints, but neither scale supports a uniform surrogate ranking or faithful practical reconstruction.

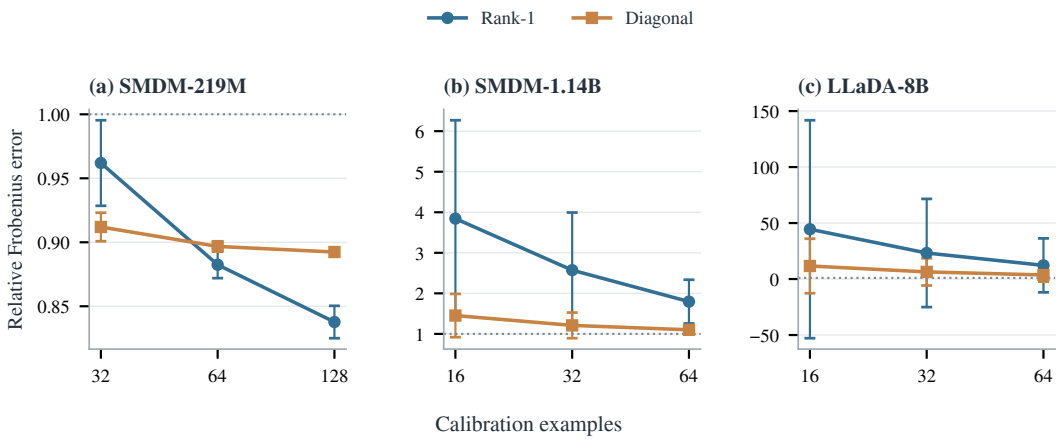


Figure 6: Held-out direct error versus calibration size. Each point first averages the fixed layer–mask grid within a seed; error bars are one sample SD across seed means.

Table 5: Complete nested-prefix audit. Values are mean±sample SD across test-seed grid means; win rate is over test-seed–layer–mask cells.

model	n_t	**rank-1**	**diagonal**	Δ	**rank-1 win rate**
SMDM-219M	16	$.969 \pm .021$	$.962 \pm .005$	$-.007 \pm .016$	29.3%
	32	$.927 \pm .020$	$.942 \pm .008$	$+.015 \pm .013$	34.7%
	64	$.907 \pm .016$	$.923 \pm .008$	$+.016 \pm .011$	34.7%
	128	$.892 \pm .019$	$.899 \pm .004$	$+.008 \pm .016$	32.0%
	256	$.869 \pm .014$	$.877 \pm .005$	$+.007 \pm .011$	32.0%
SMDM-1.14B	16	33.65 ± 7.92	13.00 ± 2.86	-20.66 ± 5.06	33.3%
	32	23.15 ± 23.76	9.05 ± 8.70	-14.11 ± 15.06	33.3%
	64	40.86 ± 42.03	15.57 ± 15.44	-25.30 ± 26.59	35.6%
	128	21.43 ± 17.94	8.38 ± 6.57	-13.06 ± 11.37	40.0%
	256	12.13 ± 8.32	4.93 ± 3.05	-7.20 ± 5.27	33.3%
LLaDA-8B	16	$1.318 \pm .209$	$1.293 \pm .286$	$-.025 \pm .148$	28.9%
	32	$1.343 \pm .229$	$1.147 \pm .207$	$-.196 \pm .037$	17.8%
	64	$1.110 \pm .134$	$.870 \pm .117$	$-.240 \pm .027$	11.1%
	128	$1.072 \pm .102$	$.841 \pm .075$	$-.231 \pm .036$	8.9%
	256	$.955 \pm .065$	$.724 \pm .052$	$-.231 \pm .018$	13.3%

B.4 Cross-Text Controls

To broaden the real-gradient inputs beyond GSM8K, we independently recompute the complete audit-qualified MT-Bench (Zheng et al., 2023) and reversal (Berglund et al., 2023) controls produced in the audit pipeline of a separately scoped concurrent study (Anonymous Authors, 2027). These are reused audit artifacts, not new runs for this paper or results reported in the companion submission. They use SMDM-219M normalization slices, and the two sources differ in both document semantics and preprocessing; they therefore test source-plus-pipeline sensitivity, not a causal corpus-content intervention.

The geometric-mean rank-1/diagonal ratios are 0.985 and 0.970, yet rank-1 wins only 17/45 and 16/45 raw seed-cells, respectively. The additional text pipelines therefore preserve the same qualitative message as GSM8K: small aggregate differences coexist with substantial cell-level heterogeneity.

B.5 Continual-Learning Context

We reanalyze the complete audit-qualified 33-run continual matrix from the separately scoped concurrent continual study (Anonymous Authors, 2027). This is an end-to-end training control

Table 6: Per-seed common-update directional fidelity.

	r_δ		q_δ	
seed	**rank-1**	**diag.**	**rank-1**	**diag.**
3407	.9986	.9998	.0014	.0002
3408	.7743	.9948	.2257	.0052
3409	.5106	.9968	.4894	.0032

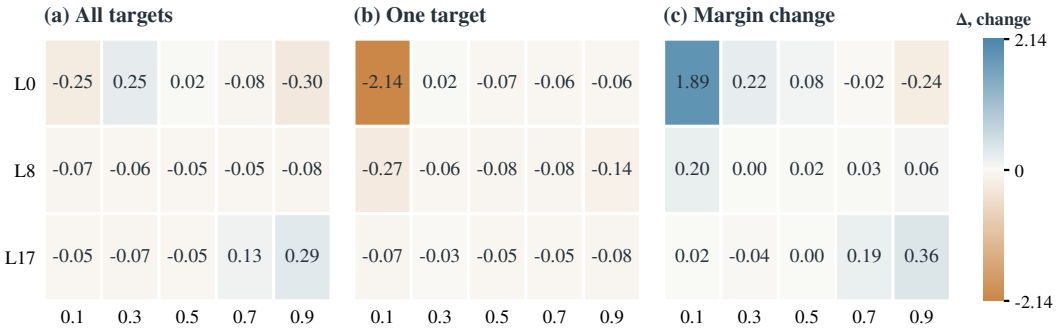


Figure 7: **Paired target-count intervention on SMDM-219M.** The first two panels use identical examples, masks, and corrupted contexts, supervising all corrupted targets or one selected target. The right panel is the paired change in $\Delta = e_{\text{diag}} - e_{\text{rank1}}$; positive values indicate a shift toward rank-1 relative to diagonal, not necessarily a rank-1 advantage. Horizontal axis: context mask probability; vertical axis: layer index.

rather than a new experiment for this paper: full-parameter SMDM-219M learns four synthetic factual-association tasks for 1,000 steps per task, in both orders and over three optimization seeds. Prompt-conditioned generative distillation (GD), rank-1 EWC, and diagonal EWC are evaluated with native random masks and independent current/replay random streams. The endpoints in Table 9 are teacher-forced losses; lower is better.

GD supplies the repeatable large improvement over sequential training. Rank-1+GD minus GD in final loss is -0.132 ± 0.603 forward and -0.285 ± 0.214 reverse, with 2/3 paired wins in either order. Relative to Diagonal+GD, the corresponding differences are $+0.145 \pm 0.314$ and -0.176 ± 0.159, again with 2/3 wins. Thus this matrix does not establish a stable incremental benefit for rank-1. All exact-match rates are zero, and the selected EWC strengths and substantial clipping further limit external validity. The result supplies bounded evidence that downstream utility and Frobenius fidelity can differ; it does not link the geometry measurements above causally to forgetting.

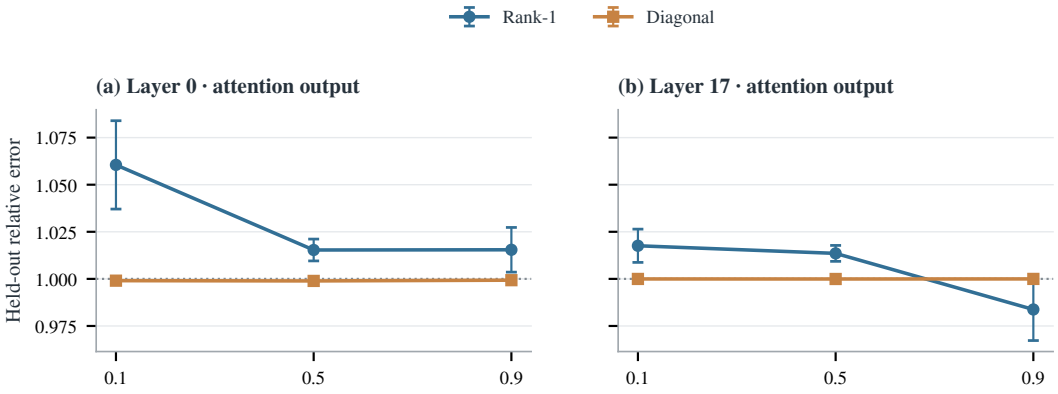


Figure 8: **Mask probability.** Dense attention-output controls in SMDM-219M. Points and bars are mean±sample SD over three split/mask seeds. Diagonal has lower error in five of six measured layer–mask cells.

Table 7: SMDM-1.14B dense-slice held-out relative Frobenius error. Values are mean±sample SD over three seeds. Oracle is the best rank-1 fit to each test Fisher and is not an available estimator.

layer	mask p	rank-1	diagonal	oracle
0	0.1	2.224 ± 1.934	1.001 ± 0.002	0.230 ± 0.180
0	0.5	1.031 ± 0.099	1.000 ± 0.000	0.697 ± 0.040
0	0.9	1.069 ± 0.035	1.000 ± 0.000	0.646 ± 0.066
19	0.1	1.203 ± 0.189	1.000 ± 0.000	0.848 ± 0.064
19	0.5	0.875 ± 0.076	1.000 ± 0.000	0.546 ± 0.066
19	0.9	0.941 ± 0.032	1.000 ± 0.000	0.691 ± 0.084

Table 8: Cross-text held-out relative Frobenius error. Values are mean±sample SD over three seed grid means; the final column counts layer–mask cell means with lower rank-1 error.

text pipeline	rank-1	diagonal	$e_{\text{diag}} - e_{\text{rank1}}$	rank-1 cells
MT-Bench	0.911 ± 0.006	0.918 ± 0.002	$+0.007 \pm 0.006$	7/15
Reversal	0.895 ± 0.008	0.911 ± 0.003	$+0.016 \pm 0.011$	5/15

Table 9: Independently recomputed continual-learning endpoints, reported as mean±SEM over three seeds. Forgetting averages the increase in past-task loss; joint training has no sequential forgetting endpoint.

order	method	final loss	forgetting
Forward	Sequential	2.927 ± 0.231	2.591 ± 0.429
Forward	GD	1.114 ± 0.374	0.674 ± 0.387
Forward	Rank-1	2.189 ± 0.408	1.942 ± 0.460
Forward	Diagonal	3.131 ± 0.282	2.793 ± 0.356
Forward	Rank-1 + GD	0.982 ± 0.251	0.194 ± 0.111
Forward	Diagonal + GD	0.837 ± 0.137	0.441 ± 0.206
Forward	Joint	0.805 ± 0.193	—
Reverse	Sequential	2.168 ± 0.390	1.997 ± 0.395
Reverse	GD	0.891 ± 0.215	0.505 ± 0.106
Reverse	Rank-1 + GD	0.606 ± 0.133	0.289 ± 0.078
Reverse	Diagonal + GD	0.782 ± 0.172	0.412 ± 0.111

B.6 Matched Slice-Level Fidelity–Forgetting Intervention

We next test the fidelity–forgetting link directly within one controlled continual setup. Full-parameter SMDM-219M first learns `d2p_8-11`, after which both surrogates are fit on the same 40 designated training rows and scored on 40 disjoint test prompts with independent native mask draws. We measure six natural slices: the attention module with `norm_1` and the MLP with `norm_2` in layers 0, 8, and 17. For each slice and each of three seeds, Rank-1+GD and Diagonal+GD branch from the same post-task checkpoint and share the frozen teacher, 64-example balanced replay cache, later-task minibatches, and mask stream while learning `p2d_12-15`. All parameters remain trainable, but the Fisher penalty is restricted to the measured slice. We fix the diagonal multiplier at 1,000 and choose the rank-1 multiplier so that the two represented matrices have equal weighted trace within that slice.

Let $s = \log(e_{\text{diag}}/e_{\text{rank1}})$ denote held-out fidelity preference. Forgetting f is the increase in Task-A teacher-forced loss from the post-Task-A checkpoint to the end of Task B, and $\Delta_f = f_{\text{diag}} - f_{\text{rank1}}$ denotes forgetting preference; positive values favor rank-1 in both cases. The primary association is the Spearman correlation across all 18 seed–slice cells. Its permutation test shuffles Δ_f among slices within each seed for 100,000 fixed draws.

Table 10: Matched slice-level fidelity and forgetting. Values are mean±sample SD over three seeds. The score column reports $10^9 s$; positive Δ_f and a win indicate lower forgetting for Rank-1+GD.

slice	$10^9 s$	Δ_f	**rank-1 wins**
Layer 0 attention	6.153 ± 8.533	-0.068 ± 0.187	2/3
Layer 0 MLP	1.153 ± 1.553	$+0.228 \pm 0.287$	2/3
Layer 8 attention	3.699 ± 6.028	-0.018 ± 0.316	2/3
Layer 8 MLP	1.075 ± 1.642	$+0.115 \pm 0.115$	3/3
Layer 17 attention	0.645 ± 0.590	$+0.294 \pm 0.251$	3/3
Layer 17 MLP	2.732 ± 3.793	-0.099 ± 0.350	2/3

We do not detect the proposed association: Spearman $\rho = -0.046$, Pearson $r = 0.002$, and the within-seed permutation test gives $p = 0.899$ (two-sided). Rank-1+GD has lower forgetting in 14/18 cells and its mean paired forgetting advantage is 0.075, but this method-level tendency does not track the fidelity ranking; it also has lower final average loss in only 10/18 cells. More fundamentally, both held-out reconstruction errors are nearly one in every cell: rank-1 ranges from 0.999999982 to 1.000000000 and diagonal from 0.999999998 to 1.000000000, so s lies only between 1.25×10^{-11} and 1.60×10^{-8}. This restricted fidelity range precludes a strong claim that no relationship can exist. Under this matched model, task, and replay setting, however, slice-level Frobenius fidelity does not explain the observed variation in forgetting.

B.7 GSM8K Retention

Finally, to complement the frozen-gradient measurements with an end-to-end behavioral check, we start from the released SMDM-1.14B GSM8K checkpoint (Nie et al., 2025a; Cobbe et al., 2021). Its released two-pass decoder obtains 784/1,319 (59.44%) strict exact match in our independent full-test validation, close to the published 58.5%. We then adapt every parameter for 1,000 updates on each of two prompt-disjoint Dolly streams—summarization followed by creative writing—with separate 120-example training and 40-example held-out sets (Databricks, 2023). Dolly held-out loss and answer-token accuracy serve only as plasticity controls; we do not treat them as benchmark scores.

The three-seed comparison includes sequential adaptation, generated distillation (GD), mean-gradient rank-1 EWC plus the identical GD cache, and trace-matched diagonal EWC plus that cache. Fisher moments use 40 held-out GSM8K source problems, represented by the released recipe's two conditional training instances per problem (Deng et al., 2023). Before examining a retention endpoint, a held-out later-task-loss rule selects the largest common trace budget that keeps both EWC variants within 10% of the 100-step GD loss; this gives $c = 10^{-5}$ per parameter. The primary endpoint is strict exact match after both later tasks on all 1,319 GSM8K test questions. We additionally retain a fixed-question paired audit: conditional retention is computed only over the initially correct items in a fixed 128-question subset. Delimiter-free numeric agreement and the rate of emitting the required `####` marker are diagnostics, not substitutes for strict exact match.

Table 11: GSM8K behavioral retention from the released SMDM-1.14B math checkpoint. Full-test endpoints are mean±sample SD over three optimization seeds on all 1,319 test questions. Conditional retention is measured on the fixed 128-question audit subset that the released checkpoint initially answered correctly. Rank-only and two-pass replay rows are 128-question mechanism checks; the latter uses seed 3407. Higher exact match and conditional retention are better; lower later-task loss is better. The EWC term is the mean weighted penalty observed during the two adaptation stages.

Retention endpoints

Method	Replay	EM (%)	Δ ckpt. (pp)	Cond. ret. (%)
		Full GSM8K endpoint (1,319 questions)		
Sequential	No	22.01 ± 1.59	-37.43 ± 1.59	37.18 ± 6.41
GD	Low-cost	36.67 ± 15.51	-22.77 ± 15.51	59.83 ± 25.68
Rank-1 + GD	Low-cost	52.31 ± 2.35	-7.13 ± 2.35	86.32 ± 5.78
Diagonal + GD	Low-cost	49.23 ± 3.28	-10.21 ± 3.28	78.63 ± 4.12
		128-question mechanism checks		
Rank-1	No	56.77 ± 2.51	—	87.61 ± 3.70
		128-question released two-pass replay sensitivity		
GD	Two-pass	53.12	—	82.05
Rank-1 + GD	Two-pass	54.69	—	87.18
Diagonal + GD	Two-pass	54.69	—	85.90

Adaptation endpoints

Method	Replay	Later loss	EWC term
		Full GSM8K endpoint (1,319 questions)	
Sequential	No	7.208 ± 0.063	—
GD	Low-cost	7.249 ± 0.105	—
Rank-1 + GD	Low-cost	7.454 ± 0.064	2.9567 ± 2.2377
Diagonal + GD	Low-cost	7.239 ± 0.077	0.0020 ± 0.0003
		128-question mechanism checks	
Rank-1	No	7.713 ± 0.195	2.0655 ± 1.6406
		128-question released two-pass replay sensitivity	
GD	Two-pass	7.414	—
Rank-1 + GD	Two-pass	7.467	0.4842
Diagonal + GD	Two-pass	7.338	0.0013

The released repository provides a math-tuned checkpoint only at this scale. Our reduced-data feasibility screen also covers the 219M, 401M, 554M, and 1.14B base checkpoints, but its 5,000-update unaugmented recipe is not matched to the released 40-epoch augmented training procedure and none passes the predeclared 1% strict-accuracy gate. We therefore do not turn that screen into a cross-scale retention comparison.

On the full test set, sequential adaptation retains $22.01 \pm 1.59\%$ strict exact match and low-cost GD retains $36.67 \pm 15.51\%$. Both Fisher-regularized variants are higher: $52.31 \pm 2.35\%$ for Rank-1+GD and $49.23 \pm 3.28\%$ for Diagonal+GD. Rank-1+GD exceeds GD by 15.64 ± 13.54 percentage points in all three seeds, but the corresponding fallback-numeric margin is only 5.66 points while the delimiter-rate margin is 14.61 points. Thus much of the strict-score gap tracks preservation of the required output marker under the low-quality replay cache.

The rank-1-specific contrast is substantially weaker. Rank-1+GD minus Diagonal+GD is $+3.08 \pm 5.04$ points and changes sign across seeds (two wins, one loss). Rank-1 also has higher later-task loss (7.454 ± 0.064 versus 7.239 ± 0.077), while equal implemented trace yields very different realized weighted EWC terms (2.9567 ± 2.2377 versus 0.0020 ± 0.0003). With released two-pass replay on

the 128-question sensitivity subset, Rank-1+GD and Diagonal+GD tie at 54.69%, versus 53.12% for GD. These results support a retention benefit from Fisher regularization in this stream, but do not establish a stable advantage specific to the mean-gradient rank-1 surrogate.

C Reproducibility and Audit

The primary released manifest validates 786 split-sample rows; its reported claims use 693 fixed-grid rows (585 primary normalization-slice rows, 90 paired target-count rows, and 18 dense-slice rows). Each raw envelope records its command, software, checkpoint/data/source hashes, seed, mask setting, loss mode, and metrics. The paired target-count envelopes additionally retain source-record, mask, Gram, cross-Gram, and diagonal sufficient statistics from which the two primary errors are recomputed.

The test-size and common-update audits form a separate 14-record validation family: 11 fixed-calibration convergence records and three full-parameter update records. Its summarizer requires every seed, grid cell, and prefix; checks fixed calibration hashes, varying test masks, disjoint held-out sources, driver hashes, and identical update scope; and regenerates Tables 5 and 6. These records remain separate from the sanitized public bundles described below.

SMDM probes ran with Python 3.9, PyTorch 2.4.1+cu121, and bfloat16 forward/backward; LLaDA probes ran with Python 3.10 and PyTorch 2.5.1+cu121. In both cases logits and transferred gradient slices use float32 and the matrix-free Gram/reconstruction algebra uses float64. The manifest checks complete Cartesian grids, contiguous seeds, finite metrics, disjoint-test metadata, oracle inequalities, and content hashes for current probes, model sources, data, checkpoints, raw evidence, and figure data.

From the repository root, the public gzip envelopes are under `iclr_1/evidence/raw/runs/`; figure values and input hashes are in `assets/iclr_1/figures/figure_data.json`. Acquisition and exact commands are documented beside the manuscript and in the code release. The release verifier checks every compressed hash and rejects identity-bearing paths; each sanitized envelope retains a content hash for its internal source artifact.

The exploratory extensions in Appendix Section B are kept separate from the 786-row primary release manifest. Their machine-readable summary, three complete anonymized 1.14B dense envelopes, validation report, and independently verified manifest are under `iclr_1/evidence/r18_completion/`. The same summarizer verifies release hashes and independently recomputes every reused cross-text and continual-learning endpoint from raw envelopes. The 100-step development smoke run is excluded from this bundle and contributes no reported number.

Invalidated runs (a Frobenius-norm bug, same-sample analyses, and unmatched pilots) are excluded; their rationale is retained in `iclr_1/report/INVALIDATED_RESULTS.md`.

The MNIST source-setting check uses the released source-paper implementation at commit `c7577f2`, Python 3.10, PyTorch 2.7.0+cu128, and three independent training seeds. Each audit records the trained-checkpoint hash, the exact calibration and test indices, independent noise seeds, software, command, and source hashes. The matrix-free computation uses float64 sample-Gram statistics and never materializes the parameter-space Fisher. The companion summarizer requires all 27 seed–timestep cells, verifies the disjoint splits and oracle ordering, and regenerates Table 4 and Fig. 5 from the three complete audits. Their identity-scrubbed envelopes and independently verified manifest are under `iclr_1/evidence/r19_mnist_unet/`, separate from the primary and R18 manifests.