TWO CHECKPOINTS ARE ENOUGH: DATA-FREE POST-HOC REPAIR OF CATASTROPHIC FORGETTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning a language model for a target task degrades capabilities that the training data never explicitly threatened, yet it also improves other held-out skills as a useful side effect. Low-rank fine-tuning methods limit the forgetting but forfeit these incidental gains. Furthermore, recovering the damaged capabilities usually requires the data the model was originally trained on, which practitioners who fine-tune open-weight models rarely have. We study post-hoc repair of this catastrophic forgetting using only the pre-fine-tuned checkpoint W_{base} and its fine-tuned descendant W_{ft}. The goal is to recover the damaged capabilities while preserving the target-task gains and incidental held-out improvements that reversion toward the base discards. We show that a single fine-tuning update separates in singular-value space into a small set of task-aligned spikes and a broad bulk that matches the Marchenko-Pastur prediction for an IID noise matrix. DG-Hard applies the Gavish-Donoho optimal hard singular-value threshold to each weight-delta matrix, keeps the spikes, and reverts the bulk. The cutoff is closed-form, so the repair needs no further training, gradients, data, or hyperparameters. To evaluate such methods, we also introduce a partition-conditional metric that separately tracks healing, preservation, non-damage, and target-task retention. Across nine held-out benchmarks, DG-Hard attains both the highest combined score and the highest mean held-out accuracy among non-spectral training-time and post-hoc baselines. We show that the retained signal carries the incidental held-out gains, and a model rebuilt from the discarded bulk alone keeps only a third of the incidental held-out gains. Alternative truncation rules such as fixed-rank and fixed-energy truncation reach or exceed DG-Hard only when their retained fraction is tuned against old-task performance, while DG-Hard's closed-form cutoff needs no tuning. The same repair also recovers most of the safety alignment eroded by benign fine-tuning, without alignment data. The code is available at `https://anonymous.4open.science/r/dghard-repair-2D84/`.

1 INTRODUCTION

Fine-tuning can improve a model on the task it is trained for while destroying capabilities that were already present in the pretrained checkpoint. In our experiments, fine-tuning Qwen3.5-4B on the medical domain raises medical-question-answering accuracy but drastically reduces math-reasoning accuracy. This failure mode is not isolated. Across the 14 model and target-task combinations we evaluate, 13 lose at least 3 pp on at least one held-out benchmark, two of them catastrophically (Tab. 11, App. E). Such failures are the modern large-model form of *catastrophic forgetting* (McCloskey & Cohen, 1989), in which adaptation to a new objective can overwrite, distort, or suppress behavior the base model had already acquired (Ratcliff, 1990; French, 1999).

Fine-tuning is the standard way pretrained models are specialized (Brown et al., 2020; Howard & Ruder, 2018; Devlin et al., 2019; Ouyang et al., 2022), yet its objective contains no term for preserving unrelated capabilities. This omission has documented consequences for general-knowledge accuracy (Luo et al., 2025), safety alignment (Qi et al., 2024), and the geometry of pretrained features outside the training-data span (Kumar et al., 2022). The checkpoint moves in two ways at once, through a structured update lowering the target-task loss and through a residual from many mini-batch SGD steps whose noise scale depends on batch size (Jastrzębski et al., 2018; Keskar et al., 2017).

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.

The loss rewards only the first. The second spreads across every direction, leaving the fine-tuned checkpoint a mixture of task-aligned signal and an unstructured residual.

We study the post-hoc repair of forgetting that has already happened. Given only a base checkpoint W_{base} and a fine-tuned checkpoint W_{ft}, the goal is to recover damaged held-out capabilities while preserving the intended and incidental gains of fine-tuning. Reversion toward the base recovers forgotten behavior cheaply but also removes the task-aligned update. A useful repair must distinguish the part of the fine-tuning delta $\Delta = W_{\text{ft}} - W_{\text{base}}$ that carries the new task from the residual.

The post-hoc repair baselines we compare against make this decision in coordinate space, by scalar interpolation (WiSE-FT, Wortsman et al., 2022), random dropping (DARE, Yu et al., 2024), magnitude- and sign-aware pruning (TIES, Yadav et al., 2023), or forgetting-aware pruning (FAPM, Huang et al., 2025). They all face the same representation problem, since task-relevant and residual entries are interleaved by magnitude, sign, and position.

The same update is more separable in singular-value space. Across fine-tuning deltas, the spectrum splits into a broad bulk that matches the random-matrix prediction (Marčenko & Pastur, 1967) and a smaller number of outlying spikes that carry the task-aligned update. Both halves have prior support, from rank-deficient updates and low intrinsic dimension (Hu et al., 2022; Aghajanyan et al., 2021) and from Marchenko-Pastur bulks in trained weights (Thamm et al., 2022; Staats et al., 2023). App. A checks both on our own deltas. Repair becomes matrix denoising, which keeps the structure and reverts the bulk.

We instantiate this with **DG-Hard**, a closed-form spectral repair that hard-thresholds the SVD of each fine-tuning delta matrix at the Gavish-Donoho cut (Gavish & Donoho, 2014). The surviving singular components form a spectrally pruned delta Δ^* that yields the repaired checkpoint $W^* = W_{\text{base}} + \Delta^*$ (Alg. 1). The method is data-free, gradient-free, and training-free and runs in minutes on a single GPU.

Contributions.

1. We formulate post-hoc repair of catastrophic forgetting as a recovery-preservation problem: recover damaged held-out capabilities from a fine-tuned checkpoint while preserving target-task gains and incidental held-out improvements.
2. We show that a single full-parameter fine-tuning update separates in singular-value space into a small set of task-aligned spikes and a bulk that matches the Marchenko-Pastur prediction for an IID noise matrix (App. A). Truncated SVD of a task vector is prior work (TSV-C). Our contribution is specifically the closed-form, per-matrix Gavish-Donoho cutoff (Gavish & Donoho, 2014) that separates the two with no data and no tuned rank, and the finding that, on full-parameter updates of two LLMs, reverting the bulk recovers most of the damaged held-out performance while keeping most of the fine-tuning gains. We call this repair **DG-Hard**.
3. We introduce a partition-conditional evaluation that reports recovery on damaged measurements, non-damage on unchanged ones, preservation on improved measurements, and retention on the target task. Across 14 model and target-task combinations and 9 held-out benchmarks, DG-Hard attains the highest Combined score among non-spectral baselines (Tab. 1). Tuned truncation rules reach or exceed that score only after their retained fraction is selected against old-task performance (Tab. 14, App. F).

2 Background and related work

Catastrophic forgetting. We build on the connectionist account of McCloskey & Cohen (1989) that distributed representations make any helpful update perturb the weights encoding prior tasks (Ratcliff, 1990; French, 1999). Kumar et al. (2022) formalize the modern instance in an overparameterized linear setting, and Luo et al. (2025); Qi et al. (2024) document its empirical reproduction in contemporary LLMs across general-knowledge and safety benchmarks.

Training-time approaches. Training-time methods modify the optimization step itself. Parameter-movement regularizers penalize updates weighted by per-parameter importance, estimated from the

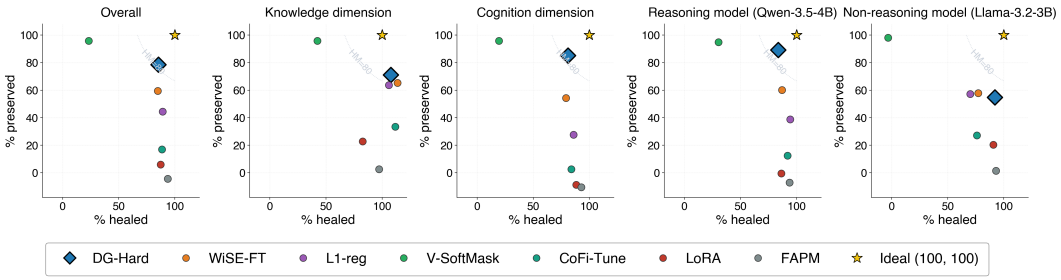

Figure 1: Recovery × preservation per cohort. Each panel plots the % healed score on the damaged partition (x-axis) against the % preserved score on the improved partition (y-axis), as defined in §3.4. The ideal corner is $(100, 100)$, and the dotted contour marks $\mathrm{HM}(\%\text{ healed}, \%\text{ preserved}) = 80$. Among non-spectral baselines, DG-Hard (blue diamond) is closest to the ideal corner across all five cohorts. FAPM (Huang et al., 2025) strongly recovers damaged measurements but sacrifices improved ones. V-SoftMask (Ke et al., 2023) preserves improved measurements but recovers less damage.

Fisher diagonal (**EWC**, Kirkpatrick et al., 2017), a path-integral contribution to the previous-task loss (**SI**, Zenke et al., 2017), or the gradient of the squared output norm on unlabeled data (**MAS**, Aljundi et al., 2018). We also compare against a generic **L1-reg** baseline that penalizes $\|W - W_{\text{base}}\|_1$ without task information. Gradient-masking methods slow the gradient through units that carry pretrained capability. **DAS** (Ke et al., 2023) scales each gradient by $(1 - \text{importance})$ from a dropout-KL proxy, and **CoFi-Tune** (Zhang et al., 2024a) adds a layer-range filter and a KL/dropout-robustness fine score. Replay methods mix or project against stored past-task examples (**GEM**, Lopez-Paz & Ranzato, 2017; **Experience Replay**, Chaudhry et al., 2019) and assume a representative pretraining sample, unavailable for LLMs with proprietary corpora.

Parameter-efficient fine-tuning. **LoRA** (Hu et al., 2022) constrains the fine-tuning update on each weight matrix to a low-rank factorization BA with $r \ll \min(m, n)$, leaving the base weights frozen. The deployed model still merges base + LoRA at inference, so forgetting can persist in the merged weights. Shuttleworth et al. (2025) show via direct spectral comparison that this low-rank constraint forces LoRA-FT updates to introduce "intruder" singular directions approximately orthogonal to the pretrained spectrum, and that these intruder directions causally drive forgetting (verified by post-hoc intervention on their singular values). Because the update is confined to a rank-r subspace, LoRA also produces little of the incidental improvement on unrelated held-out tasks that full-parameter fine-tuning yields (§3.4 counts 55 such improved measurements for full fine-tuning). It therefore limits collateral gains as well as collateral damage.

Post-hoc model merging. A first post-hoc family operates on $\Delta = W_{\text{ft}} - W_{\text{base}}$ via cheap, data-free transformations. Most target multi-task merging. With a single fine-tune, their aggregation steps (sign election, mean over task vectors) reduce to identity, leaving the per-vector preprocessing as a post-hoc repair on the $(W_{\text{base}}, W_{\text{ft}})$ pair. **WiSE-FT** (Wortsman et al., 2022) interpolates linearly, $W^* = (1-\alpha)W_{\text{base}} + \alpha W_{\text{ft}}$. **Task Arithmetic** (Ilharco et al., 2023) treats Δ as a task vector and studies algebraic operations on it. **TIES-Merging** (Yadav et al., 2023) trims low-magnitude entries, elects a per-parameter consensus sign across vectors, and averages the aligned entries. **DARE** (Yu et al., 2024) drops entries of Δ with probability p and rescales survivors by $1/(1-p)$. **FAPM** (Huang et al., 2025) prunes entries by a forgetting-aware score combining absolute change and a relative-change penalty against the pretrained weight. **Fisher-weighted averaging** (Matena & Raffel, 2022) weights each parameter by its empirical Fisher importance and needs a labeled sample. Per-method hyperparameters are in App. K.1.

Spectral task-vector methods. Closest to our method are post-hoc methods that work on the layer-wise SVD of Δ. **TSV-M** (Gargiulo et al., 2025) merges several task vectors by reducing interference among their singular vectors, and **KnOTS** (Stoica et al., 2025) aligns several LoRA models through the SVD before merging. Both act on the relation between multiple updates, and on

a single update TSV-M reconstructs Δ unchanged. **TSV-C** (Gargiulo et al., 2025) keeps the top n/T singular components of each layer task matrix, T being the number of merged tasks, for a headline compression to 10% of the original size. TSV-C and DG-Hard return the same truncated-SVD reconstruction once a rank is chosen (§4.4) and differ only in the rank rule. DG-Hard derives a per-matrix cutoff from a denoising model and tests whether reverting the removed component repairs forgetting. At $T = 1$ the TSV-C rule keeps everything, so §4.4 evaluates it at the 10% rank fraction of that compression, alongside 5%, 15%, and other rank rules.

Data-based repair and unlearning. A separate line of work repairs a fine-tuned model with data. Unlearning by gradient ascent (Jang et al., 2023) and NPO (Zhang et al., 2024b) optimize against a forget set, and RSS (Shubham et al., 2026) attributes trustworthiness failures of a fine-tuned LLM to detrimental fine-tuning examples by influence estimation and corrects them with a gradient step. These methods require the relevant data, gradients, and an optimization run, and therefore training-class memory. Repair from two checkpoints requires inference-class memory (App. B) and no knowledge of which capability degraded. The three properties act together, and only all three let whoever holds the two checkpoints run the repair. The settings are complementary. Data-based repair suits a failure that traces to identifiable examples when data and training compute are available, and we address the setting where neither is. We do not compare against these methods within their own setting.

Spectral shrinkage and our distinction. The spectral framing fits fine-tuning because the weight update is empirically rank-deficient (Hu et al., 2022; broader evidence in App. A), and random-matrix theory supplies the matching denoiser. The Marchenko-Pastur theorem (Marčenko & Pastur, 1967) fixes the asymptotic spectrum of an IID-noise matrix, and the Gavish-Donoho hard threshold (Gavish & Donoho, 2014) identifies the AMSE-optimal cut between signal and noise (Gavish & Donoho, 2017 extend to optimal continuous shrinkers). The post-hoc repair baselines above operate in coordinate space (TIES, DARE, FAPM), as a single scalar (WiSE-FT, Task Arithmetic), or via labeled-data Fisher weighting. Truncated SVD of a task vector is prior work (TSV-C), and §1 states what this paper adds.

3 METHOD

We propose **DG-Hard**, a post-hoc, data-free, gradient-free repair that strips the IID noise residual from a fine-tuning update by applying the optimal hard singular-value threshold of Gavish & Donoho (2014) to its SVD. The procedure takes only the base checkpoint W_{base} and the fine-tuned checkpoint W_{ft}, operates one 2D weight matrix at a time, and reduces to a closed-form threshold in the matrix shape and a single estimated noise scale.

3.1 SIGNAL-PLUS-NOISE MODEL OF THE FINE-TUNING UPDATE

For a single 2D weight matrix in $\mathbb{R}^{m\times n}$, write the fine-tuning update as $\Delta = W_{\text{ft}} - W_{\text{base}}$ and take its singular value decomposition,

$$\Delta \;=\; U\ \mathrm{diag}(\mathbf{s})\, V^{\top}, \quad s_1 \geq \cdots \geq s_p \geq 0, \quad p = \min(m, n). \tag{1}$$

We model Δ as a two-component sum,

$$\Delta \;=\; \Delta_{\text{signal}} + \Delta_{\text{noise}}, \tag{2}$$

with Δ_{signal} a task-aligned update, low-rank in practice (the cutoff sets no rank, App. M), and Δ_{noise} an IID residual of unknown scale σ. The SGD-noise mechanism in §1 is consistent with this structure, since mini-batch residuals accumulate into the weights while the gradient has no incentive to remove them. IID entries are our modeling assumption. Both components have empirical support. Fine-tuning updates are rank-deficient (Hu et al., 2022) with low intrinsic dimensionality (Aghajanyan et al., 2021), and trained-weight spectra fit the Marchenko-Pastur bulk past a finite number of outliers (Thamm et al., 2022; Staats et al., 2023). A direct check on our own deltas is consistent with both layer-locally (App. A, Fig. 3). Our goal is to remove Δ_{noise} while leaving Δ_{signal} intact.

Algorithm 1 DG-Hard repair (per weight matrix).

Require: Base weight $W_{\text{base}} \in \mathbb{R}^{m\times n}$, fine-tuned weight $W_{\text{ft}} \in \mathbb{R}^{m\times n}$.
Ensure: Repaired weight $W^* \in \mathbb{R}^{m\times n}$.
1: $\Delta \leftarrow W_{\text{ft}} - W_{\text{base}}$
2: $(U, \mathbf{s}, V) \leftarrow \mathrm{SVD}(\Delta)$
3: $p \leftarrow \min(m,n)$; $\beta \leftarrow p/\max(m,n)$
4: $\hat{\sigma} \leftarrow \mathrm{median}(\mathbf{s})/\big(\mu_\beta\sqrt{\max(m,n)}\big)$ ▷ noise scale, App. B
5: $\tau \leftarrow \lambda^*(\beta)\,\hat{\sigma}\,\sqrt{\max(m,n)}$ ▷ optimal threshold, (4)
6: $s'_i \leftarrow s_i \cdot \mathbf{1}\{s_i > \tau\}$ **for** $i = 1,\dots,p$
7: $\Delta^* \leftarrow U\ \mathrm{diag}(\mathbf{s}')\,V^\top$
8: **return** $W_{\text{base}} + \Delta^*$

3.2 Random-matrix theory and the bulk edge

If Δ were pure IID noise of variance σ^2, the Marchenko-Pastur theorem (Marčenko & Pastur, 1967) pins its spectrum exactly. In the limit $m, n \to \infty$ at fixed aspect ratio $\beta = \min(m,n)/\max(m,n)$, every singular value sits below the bulk edge

$$\lambda_{\text{MP}} \;=\; \sigma\left(1+\sqrt{\beta}\right)\sqrt{\max(m,n)} \tag{3}$$

with probability one, depending only on σ and β. In that limit a singular value $s_i > \lambda_{\text{MP}}$ of Δ is not produced by the noise component of (2). The edge is a principled boundary, but denoising requires a *shrinker* mapping each s_i to a denoised s'_i. The simplest principled choice is a hard threshold. Gavish & Donoho (2014) derive the one that minimizes the asymptotic mean squared error (AMSE), the $m, n \to \infty$ limit at fixed β of the expected reconstruction error $\mathbb{E}\big[\|\hat{\Delta}_{\text{signal}} - \Delta_{\text{signal}}\|_F^2\big]$ (the squared Frobenius distance between the estimated and true low-rank signal). A low threshold admits noise (high variance), a high threshold discards signal (high bias), and the AMSE-optimal threshold sits at the trade-off minimum.

3.3 DG-Hard: the optimal hard threshold

Among all hard shrinkers $\eta(s) = s \cdot \mathbf{1}\{s > \tau\}$, Gavish & Donoho (2014) identify the threshold minimizing the AMSE of the reconstruction under (2):

$$\tau^* = \lambda^*(\beta)\,\sigma\,\sqrt{\max(m,n)}, \qquad \lambda^*(\beta) = \sqrt{2(\beta+1) + \frac{8\beta}{(\beta+1)+\sqrt{\beta^2+14\beta+1}}}\,. \tag{4}$$

For square matrices, $\lambda^*(1) = 4/\sqrt{3} \approx 2.309$, the constant for which the original paper is named. The DG-Hard shrinker is

$$\eta_{\text{DG}}(s) \;=\; s \cdot \mathbf{1}\big\{\, s > \tau^* \big\}. \tag{5}$$

τ^* sits strictly above λ_{MP} ($\approx\ 2.309\,\sigma\sqrt{n}$ against $2\sigma\sqrt{n}$ at $\beta = 1$), since a cut at the edge admits singular values barely distinguishable from noise at full variance cost. Eq. (4) is the unique point at which marginal bias and variance balance. Algorithm 1 gives the full per-matrix procedure. The noise estimator $\hat{\sigma}$ on line 4 is the median-based Gavish-Donoho estimator $\hat{\sigma} = \mathrm{median}(\mathbf{s})/(\mu_\beta\sqrt{\max(m,n)})$, which substitutes into (4) to give a data-only threshold $\tau^* = \omega(\beta)\ \mathrm{median}(\mathbf{s})$ with $\omega(\beta) = \lambda^*(\beta)/\mu_\beta \approx 2.858$ for square matrices, the unknown-noise coefficient of Gavish & Donoho (2014). App. B gives the full derivation and the SVD-dominated cost ($O(\min(m,n)^2\max(m,n))$ per matrix).

3.4 Evaluation protocol

For every held-out benchmark b, cell, and method we read the accuracy score_b (in pp) and define

$$\Delta_{\text{FT}}(b) = \mathrm{score}_{\text{FT}}(b) - \mathrm{score}_{\text{base}}(b), \qquad \Delta_{\text{method}}(b) = \mathrm{score}_{\text{method}}(b) - \mathrm{score}_{\text{base}}(b). \tag{6}$$

At a ± 3 pp threshold, each (model, task, b) triple is **damaged** if $\Delta_{\text{FT}}(b) \le -3$, **improved** if $\Delta_{\text{FT}}(b) \ge +3$, and **unchanged** otherwise. Varying the threshold from 2 to 6 pp keeps DG-Hard first among non-spectral baselines (App. C). The partition is fixed by the FT checkpoint and identical

across methods, so per-method comparisons use the same set of triples in each partition class. Across the 126-triple matrix (2 models $\times$ 7 tasks $\times$ 9 benchmarks) we observe 30 damaged, 55 improved, 41 unchanged triples.

Per-method statistics. Let $s_M = \text{score}_{\text{method}}$, $s_F = \text{score}_{\text{FT}}$, and $s_B = \text{score}_{\text{base}}$. Let D, I, U denote the damaged, improved, and unchanged partition sets, and HM the harmonic mean. Each method produces seven percentage scores. Higher is better, and 100 marks full recovery or preservation. Values above 100 indicate overshoot, values below 0 indicate regression, and HM inputs are floored at 0. % healed and % preserved are ratios of sums, so each triple is weighted by the size of its fine-tuning effect (App. C discusses the mean-of-ratios alternative):

$$\textbf{\% healed} = \frac{\sum_{b\in D}\big(s_M(b) - s_F(b)\big)}{\sum_{b\in D}\big(s_B(b) - s_F(b)\big)} \cdot 100, \quad \textbf{\% preserved} = \frac{\sum_{b\in I}\big(s_M(b) - s_B(b)\big)}{\sum_{b\in I}\big(s_F(b) - s_B(b)\big)} \cdot 100, \tag{7a}$$

$$\textbf{on-task ret.} = \operatorname*{avg}_{\text{cells}} \frac{s_M(\text{task})}{s_F(\text{task})} \cdot 100, \quad \textbf{non-damage} = \operatorname*{avg}_{b\in U} \mathbf{1}\{s_B(b) - s_M(b) < 3\} \cdot 100, \tag{7b}$$

$$\textbf{Clean-up} = \text{HM}(\%\text{ healed, non-damage}), \quad \textbf{Retention} = \text{HM}(\%\text{ preserved, on-task ret.}), \tag{7c}$$

$$\textbf{Combined} = \text{HM}(\text{Clean-up, Retention}). \tag{7d}$$

% healed is the share of the accuracy lost on D that is recovered, **% preserved** the share of the accuracy gained on I that is kept, **on-task ret.** the on-task accuracy held vs. FT, and **non-damage** the share of unchanged triples less than 3 pp below base. "Combined" is the ranking statistic. It is a nested harmonic mean, so a collapse in any of the four sub-statistics drives it toward 0, demoting methods that win one axis by sacrificing another (V-SoftMask preserves without healing, and FAPM heals without preserving). DG-Hard remains first among non-spectral methods under the alternative aggregators and the mean-of-ratios variant. App. C contrasts the metric with flat unified scores (Zhang et al., 2024a).

4 Experiments

A *cell* is one of the 14 fine-tuned (model, task) checkpoints, and a *triple* is one (cell, held-out benchmark) pair. The 14 cells expand into a 126-triple matrix. The headline finding across this matrix is that DG-Hard attains the highest Combined score of the non-spectral methods, overall and on each model. On Llama its lead over WiSE-FT is within the bootstrap interval, and WiSE-FT wins more of the individual Llama cells (§4.2). Tuned spectral truncation matches DG-Hard (§4.4). Full per-method per-cell scores are in App. K.

4.1 Setup at a glance

Two models span the reasoning / non-reasoning split, **Qwen3.5-4B** (Qwen Team, 2026) (thinking mode, `<think>` blocks stripped at scoring) and **Llama-3.2-3B-Instruct** (Grattafiori et al., 2024; Meta, 2024). Each is fine-tuned under one full-parameter SFT configuration on seven tasks (RTE, StrategyQA, ReClor, BoolQ, MedQA, WikiQA, Winogrande) and evaluated on nine held-out benchmarks: a *Knowledge* cohort (MMLU, TriviaQA, TruthfulQA) and a *Cognition* cohort (ARC-Challenge, GSM8K, IFEval, Math-500, MNLI, HellaSwag). The nine compared methods fall into three groups: reference points **Base** (post-alignment, unfine-tuned) and **Full-SFT** (unrepaired fine-tuned); training-time interventions **L1-reg**, **V-SoftMask** (DAS, Ke et al., 2023), **CoFi-Tune** (Zhang et al., 2024a), and **LoRA** (Hu et al., 2022) (rank 16); and post-hoc methods **WiSE-FT** (Wortsman et al., 2022), **FAPM** (Huang et al., 2025), and our **DG-Hard**. Spectral controls are evaluated in §4.4. Element-wise merging baselines (TIES (Yadav et al., 2023), DARE (Yu et al., 2024), Task Arithmetic (Ilharco et al., 2023)) appear in App. L.3. Inference is identical across methods (vLLM, greedy decoding, no `max_tokens` cap), with a measured ± 1 pp drift floor and a 3 pp significance threshold ($\approx 3\times$ the drift floor). Identifiers, splits, hyperparameters, repair scope, method configurations, and inference details are in App. D, K.1, and K.2.

Table 1: Population-level Clean-up, Retention, and Combined scores per method. Among these methods, DG-Hard is the only one scoring ≥ 80 on both Clean-up and Retention. WiSE-FT (Wortsman et al., 2022) comes closest but falls to 73.8 on Retention. Every other baseline falls to 61.3 or lower on one of the two axes. The last two columns report the traditional averages over the 14 cells: mean accuracy across the nine held-out benchmarks, and mean on-task accuracy. Base and Full-SFT are reference rows. **Bold** marks the column maximum and any value within its reported error. Our method's row is shaded.

Method	% healed ↑	Non-damage ↑	% preserved ↑	On-task ret. ↑	Clean-up ↑	Retention ↑	Combined ↑	Mean held-out ↑	Mean on-task ↑
Base	–	–	–	–	–	–	–	60.0	–
Full-SFT	–	–	–	–	–	–	–	63.3	**83.7**
DG-Hard	$85.4_{\pm 1.3}$	$85.4_{\pm 2.0}$	$78.5_{\pm 0.6}$	$97.3_{\pm 0.2}$	$85.4_{\pm 1.4}$	$86.9_{\pm 0.4}$	**86.1**	**64.4**	81.6
WiSE-FT	$84.8_{\pm 1.2}$	$90.2_{\pm 1.8}$	$59.4_{\pm 0.6}$	$97.5_{\pm 0.2}$	$87.5_{\pm 1.2}$	$73.8_{\pm 0.5}$	80.1	63.2	81.6
L1-reg	$89.1_{\pm 1.2}$	$85.4_{\pm 2.0}$	$44.5_{\pm 0.7}$	$98.6_{\pm 0.2}$	$87.2_{\pm 1.4}$	$61.3_{\pm 0.7}$	72.0	62.4	82.6
CoFi-Tune	$88.5_{\pm 1.3}$	$75.6_{\pm 2.1}$	$17.0_{\pm 1.0}$	$96.2_{\pm 0.2}$	$81.5_{\pm 1.5}$	$28.9_{\pm 1.5}$	42.7	60.6	80.7
V-SoftMask	$23.2_{\pm 0.9}$	$\mathbf{97.6_{\pm 1.9}}$	$\mathbf{95.8_{\pm 0.6}}$	$\mathbf{99.6_{\pm 0.2}}$	$37.5_{\pm 1.1}$	$\mathbf{97.7_{\pm 0.3}}$	54.2	63.8	83.4
LoRA	$87.3_{\pm 1.2}$	$85.4_{\pm 2.0}$	$6.0_{\pm 1.0}$	$94.9_{\pm 0.2}$	$86.3_{\pm 1.4}$	$11.3_{\pm 1.7}$	20.0	59.6	79.7
FAPM	$\mathbf{93.6_{\pm 1.3}}$	$\mathbf{97.6_{\pm 1.7}}$	$-4.4_{\pm 1.1}$	$77.8_{\pm 0.2}$	$\mathbf{95.5_{\pm 1.3}}$	$0.0_{\pm 0.0}$	0.0	59.5	64.8

4.2 Per-cell forgetting and repair winners

Fine-tuning damages at least one held-out benchmark in 13 of 14 cells, in two of them catastrophically (Qwen/MedQA's GSM8K collapses from 92.9% to 1.1%). Damage concentrates on a few benchmarks per cell, and the matrix holds 55 improved triples against 30 damaged ones. A cell-level mean held-out score therefore smooths the catastrophic drops away and is the wrong instrument for detecting forgetting, so the section evaluates methods per benchmark (per-cell breakdown in Tab. 11, App. E). After DG-Hard repair, damaged triples fall from 30 to 17 (10 of the original 30 remain and 7 are new, the largest 18.8 pp). The on-task cost averages 2.2 pp (Tab. 12, App. E).

A per-cell balance score, HM of the method's mean held-out ratio against base and its on-task ratio against Full-SFT ($\times 100$), compares DG-Hard with WiSE-FT, the strongest non-spectral baseline. **DG-Hard wins 6 of 7 Qwen cells (plus 1 tie). WiSE-FT wins 4 of 7 Llama cells (plus 2 ties). Of the 14 cells, DG-Hard wins 7, WiSE-FT 4, and 3 tie** (Tab. 13, App. E).

4.3 Recovery vs preservation (population-level)

Recovery and preservation are coupled. Reverting W_{ft} toward W_{base} trivially maximizes recovery but discards the held-out and on-task gains fine-tuning produced. Tab. 1 reports the population-level Clean-up, Retention, and Combined scores (Eq. 7, §3.4) for every method. On the traditional averages (last two columns) DG-Hard also has the highest mean held-out accuracy, 64.4 against 63.3 for Full-SFT and 60.0 for Base, at an on-task cost of 2.2 pp (Tab. 12). The means sit close together because each pools 30 damaged triples with 55 improved ones, and the partition-conditional scores separate them.

FAPM (95.5/0.0) and V-SoftMask (37.5/97.7) are the one-axis extremes. LoRA (86.3/11.3) cleans up well but has no FT-side gain to preserve (§2). CoFi-Tune, L1-reg, and WiSE-FT clean up well and retain less.

4.4 Spectral controls

DG-Hard belongs to a family of truncated-SVD edits of Δ. Once a rank is chosen for a matrix, all members return the same Eckart-Young reconstruction (Eckart & Young, 1936), and they differ only in how the rank is chosen (Tab. 14, App. F, all 14 cells). **Fixed-rank** truncation keeps a fixed fraction of the rank. The 10% setting is the TSV-C fraction (§2). **Fixed energy-percentage** keeps the fewest leading components that reach a fixed fraction of the squared-singular-value energy. **Effective-rank** keeps $k = \text{round}(\exp(H))$ components, with H the entropy of the normalized singular values (Roy & Vetterli, 2007). **DG-soft**, the continuous-shrinkage counterpart of the hard cut, applies the Frobenius-optimal shrinker of Gavish & Donoho (2017) with the same noise estimate, zeroing below the Marchenko-Pastur edge and shrinking every surviving value. **Base + noise-bulk** drops the signal and keeps the sub-threshold bulk that DG-Hard discards.

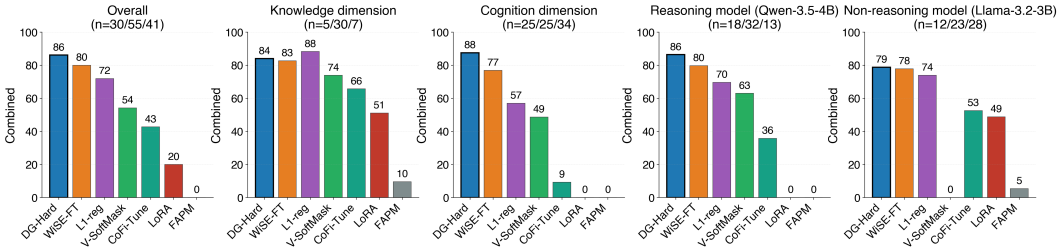

Figure 2: Combined score per method and cohort. Panel titles give the (damaged/improved/unchanged) triple counts. Among non-spectral baselines DG-Hard leads on four cohorts (§4.5). A collapse on either Clean-up or Retention pulls Combined toward zero.

Healing is cheap, keeping the gains while healing is not. Base + noise-bulk heals as much as DG-Hard (84.7% against 85.4%), and so do uniform shrinkage (WiSE-FT, 84.8%) and forgetting-aware pruning (FAPM, 93.6%), so healing follows from breaking the spike-and-bulk combination, by removing one component or attenuating both. The methods separate on what they keep. FAPM preserves none of the held-out gains, base + noise-bulk 35.5%, WiSE-FT 59.4%, and DG-Hard 78.5%, and on Cognition the bulk's Retention is 35.4 against 90.8. The spikes carry the incidental gains, and keeping them while the bulk goes is what the spectral split buys. The two components are interleaved. Either one of them, on its own, leaves the damaged benchmarks near base and keeps the target task, and the damage appears only when both are present. DG-Hard also removes more of the update than WiSE-FT (11% of the energy kept against 25%) while preserving more, and at comparable rollback it leads uniform interpolation by 22 pp and coordinate-wise merges by more (App. L.3). Keeping the spikes together with most of the bulk fails the other way. Energy-percentage at 90% and effective-rank heal only 16.1% and 10.7% of the damage, because the task signal is low-rank yet carries a minority of the energy.

The hard cutoff sets the balance. DG-soft heals slightly more than DG-Hard (88.8 against 85.4) and preserves less (Retention 78.1 against 86.9). Truncation at TSV-C's 10% fraction heals 64.0%.

Tuned truncation matches DG-Hard. Fixed-rank at 5% and energy-percentage at 15%, the best settings of their families, reach Combined 86.9 and 87.9 against 86.1 for DG-Hard, and under a paired per-prompt bootstrap the 95% interval of each difference includes zero (energy-percentage at 15% exceeds DG-Hard in 73% of resamples, fixed-rank at 5% in 53%). The retained fraction must be selected against old-task performance or against a calibration set (training-free, but each candidate costs an evaluation run), and therefore the repair is no longer data-free. DG-Hard fixes the fraction per matrix in closed form from the noise model, retaining about 3% of the rank and on average 15.5% of the energy (range 10.3 to 21.9% across the 14 checkpoints; Tab. 3, App. A), next to where the tuned rules peak (15% energy, 5% rank). Per cell, the closed-form cutoff holds up. The best rule changes across the 14 cells (seven of the eleven rules win at least one), and DG-Hard has the lowest median regret to the per-cell best rule in HM(% healed, % preserved) (3.5 points, against 5.7 for energy-percentage 15%), with its lead over energy 15% inside the noise of 14 cells (App. F). What the closed-form cutoff buys is low regret with no fraction to select or calibrate, staying close to the best tuned rule in most cells without topping every one.

4.5 Cohort breakdown

DG-Hard exceeds every non-spectral baseline on Combined on four of five cohorts (Overall, Cognition, Reasoning, Non-reasoning) by keeping Clean-up and Retention strong at once (Fig. 2; per-cohort Clean-up and Retention in App. G). On the small-n Knowledge cohort (5 damaged triples), L1-reg leads on Combined (88.3 against 84.0) through its non-damage rate (100 against 71.4). This advantage does not transfer to the larger Cognition cohort (L1-reg Retention 43.1, Combined 57.0).
DG-Hard's edge over WiSE-FT comes from Retention. WiSE-FT's Clean-up is ahead by 1.5 to 5.7 pp on four cohorts and behind by 6.4 on Non-reasoning. DG-Hard's Retention, however, is 13 pp or more higher on Overall, Cognition, and Reasoning, driven by held-out preservation (on-task

retention is $\sim 97\%$ for both). On Non-reasoning the order reverses (Retention 69.7 against 72.4) and DG-Hard's edge comes from Clean-up.

4.6 Cognition and the trade-off frontier

The Knowledge cohort ($n = 5$ damaged) is too easy for healing to differentiate methods, since DG-Hard, WiSE-FT, L1-reg, and CoFi-Tune all recover above base on its damaged cases (% healed above 100), and FAPM comes close (Fig. 1). The Cognition cohort ($n = 25$ damaged) separates the non-spectral baselines, on Retention. DG-Hard's Cognition Retention of 90.8 is 21 pp above WiSE-FT (69.7) and far above L1-reg (43.1) and CoFi-Tune (5.0). FAPM and LoRA fall to zero because their % preserved there is negative (Tab. 17). **The Cognition preservation gap separates DG-Hard from the other non-spectral baselines and is the most plausible mechanism behind its reasoning-model lead.** V-SoftMask preserves Cognition but heals only a fifth of its damage (19.5%), while DG-Hard heals 81.0% of the Cognition damage (Fig. 1).

In the trade-off frontiers of Fig. 4 (App. I), only DG-Hard among non-spectral baselines sits in the upper-right of both panels. Bucketing by FT-effect magnitude (App. I), both heal most of the damage in every bucket, and DG-Hard preserves more in every lift bucket, with the largest lead on the largest lifts ($+84.2$ against $+58.2$).

4.7 Safety alignment recovery

Both base models are safety aligned at the Instruct tier, and benign fine-tuning is not designed to touch alignment. We ask whether FT erodes safety as a side effect and whether DG-Hard restores it. Six cells (WikiQA, MedQA, and Winogrande per model) are scored on HarmBench (Mazeika et al., 2024), XSTest (Röttger et al., 2024), and StrongREJECT (Souly et al., 2024) (Tab. 21, App. J).

The signal-plus-noise model predicts that alignment, encoded in pretrained directions Base already occupies, survives a repair that removes only the IID noise residual. Fine-tuning can move a safety axis in either direction, so recovery means returning toward Base. On every Llama cell where FT lowered refusal or raised harmfulness, DG-Hard recovers most of the gap toward Base, and no cell degrades much. The one regression is Qwen+medqa, below both Base and FT on HarmBench (91.5 against 95.0 and 99.5) and StrongREJECT. In Qwen+wikiqa, FT moves the axes in opposite directions (refusal to 100%, XSTest over-refusal from 2.8 to 15.6), and DG-Hard returns both near Base. The threshold was set without alignment data, so the safety recovery follows from the same noise-residual cut that repairs capabilities.

5 Limitations and conclusion

Limitations. DG-Hard keeps every singular direction of Δ above the Gavish-Donoho cutoff and discards the rest, so a task signal spread thinly inside the noise bulk would be lost. None of the 14 measured deltas is in that regime (App. M). Fixed-rank and fixed-energy truncation reach or exceed DG-Hard once their retained fraction is selected against old-task performance or a calibration set (§4.4), and the three are statistically level, but the tuned controls are no longer data-free and an untuned fraction is a guess. The evaluation covers two models under 5B parameters, one training configuration per cell (App. D.4), and WiSE-FT at $\alpha = 0.5$ only. Due to limited computational resources, the cutoff has not been tracked across learning rates, epochs, batch sizes, or larger models. The only operational cost is the per-matrix SVD. Even at the $\sim$ 1T-parameter scale of Kimi-K2.6, the dominant matrix's SVD grows only $\sim 7\times$ over Llama-3.2-3B's embedding, the expert blocks admit parallel SVD, and a full repair completes in hours.

Conclusion. Fine-tuning weight deltas split into a low-rank task-aligned update and an IID-like noise residual, and DG-Hard strips the residual with the Gavish-Donoho hard threshold, post hoc, data-free, and gradient-free. The damage appears only when spikes and residual are combined, and the spikes carry the incidental gains. Across 14 model and task combinations and nine held-out benchmarks, DG-Hard attains the highest Combined score among non-spectral baselines, 6.1 points above WiSE-FT, and recovers most of the safety alignment that benign fine-tuning erodes, without alignment data, so both patterns appear to follow from one mechanism. Two public checkpoints suffice, so a group without training compute can run it locally.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods in the form of preparing the fine-tuning scripts and the scoring code and refactoring the evaluation and analysis code. We have not used generative AI tools to generate synthetic datasets, to develop theoretical models or conceptual frameworks, to formulate mathematical claims, to propose or refine hypotheses, to design or provide feedback on research methodology or experiments, or to interpret results, and assistance with proofs, translation, dataset cleaning and reformatting, and qualitative or thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to assist with the writing and editing of the text and to generate tables and figures from computed results. We have reviewed all AI-assisted work. All AI-assisted code, including the fine-tuning scripts, the scoring code, and the refactored evaluation and analysis code, was reviewed and tested by the authors, every reported number was recomputed from the stored evaluation outputs by scripts the authors checked, and all AI-assisted text was edited by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The repair method is fully specified by Algorithm 1 and App. B, and an anonymized implementation is linked in the abstract. App. D lists the fine-tuning tasks, held-out benchmarks, HuggingFace identifiers, evaluation splits, training hyperparameters, and the per-model repair scope; App. K.1 gives every baseline and control configuration, and App. K.2 the inference engine and decoding settings. The evaluation metric is defined in §3.4, and App. K reports every per-benchmark score of the headline matrix from which its aggregate numbers derive.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7319–7328, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.568. URL https://aclanthology.org/2021.acl-long.568/.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Computer Vision – ECCV 2018*, pp. 144–161, Cham, 2018. Springer International Publishing. doi: 10.1007/978-3-030-01219-9_9. URL https://doi.org/10.1007/978-3-030-01219-9_9.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019. URL https://arxiv.org/abs/1902.10486.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. doi: 10.48550/arXiv.1803.05457. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367. URL https://doi.org/10.1007/BF02288367.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. doi: 10.1016/S1364-6613(99)01294-2. URL https://doi.org/10.1016/S1364-6613(99)01294-2.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18695–18705, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gargiulo_Task_Singular_Vectors_Reducing_Task_Interference_in_Model_Merging_CVPR_2025_paper.html.

Matan Gavish and David L. Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014. doi: 10.1109/TIT.2014.2323359. URL https://doi.org/10.1109/TIT.2014.2323359.

Matan Gavish and David L. Donoho. Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, 63(4):2137–2152, 2017. doi: 10.1109/TIT.2017.2653801. URL https://doi.org/10.1109/TIT.2017.2653801.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370. URL https://aclanthology.org/2021.tacl-1.21/.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://iclr.cc/virtual/2021/poster/2962.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031. URL https://aclanthology.org/P18-1031/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Wei Huang, Anda Cheng, and Yinggui Wang. Mitigating catastrophic forgetting in large language models with forgetting-aware pruning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21842–21856, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1108. URL https://aclanthology.org/2025.emnlp-main.1108/.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2023. URL https://arxiv.org/abs/2212.04089.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805/.

Stanislaw Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Width of minima reached by stochastic gradient descent is influenced by learning rate to batch size ratio. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pp. 392–402, Cham, 2018. Springer International Publishing. doi: 10.1007/978-3-030-01424-7_39. URL https://doi.org/10.1007/978-3-030-01424-7_39.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421. URL https://doi.org/10.3390/app11146421.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=m_GDIItaI3o.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://doi.org/10.1073/pnas.1611835114.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems*

Principles, SOSP '23, pp. 611–626, New York, NY, USA, 2023. ACM. doi: 10.1145/3600006.3613165. URL https://arxiv.org/abs/2309.06180.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/f87522788a2be2d171666752f97ddebb-Abstract.html.

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

V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967. doi: 10.1070/SM1967v001n04ABEH001994. URL https://iopscience.iop.org/article/10.1070/SM1967v001n04ABEH001994.

Michael S. Matena and Colin A. Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17703–17716. Curran Associates, Inc., 2022. doi: 10.52202/068431-1287. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/70c26937fbf3d4600b69a129031b66ec-Abstract-Conference.html.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.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. doi: 10.1016/S0079-7421(08)60536-8. URL https://doi.org/10.1016/S0079-7421(08)60536-8.

Meta. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. Blog post, September 2024. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. Task-specific skill localization in fine-tuned language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27011–27033. PMLR, 2023. URL https://proceedings.mlr.press/v202/panigrahi23a.html.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, volume 2024, pp. 30988–31043, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/83b7da3ed13f06c13ce82235c8eedf35-Abstract-Conference.html.

Qwen Team. Qwen3.5: Towards native multimodal agents. Blog post, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Roger Ratcliff. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, 97(2):285–308, 1990. doi: 10.1037/0033-295X.97.2.285. URL https://doi.org/10.1037/0033-295X.97.2.285.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pp. 606–610, Poznan, Poland, 2007. IEEE. URL https://ieeexplore.ieee.org/document/7098875.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740, April 2020. doi: 10.1609/aaai.v34i05.6399. URL https://ojs.aaai.org/index.php/AAAI/article/view/6399.

Pratyusha Sharma, Jordan Ash, and Dipendra Kumar Misra. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. In *International Conference on Learning Representations*, pp. 17632–17651, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4c2092ec0b1370cce3fb5965ab255fae-Abstract-Conference.html.

Kumar Shubham, Nishant Sharma, Karn Tiwari, and Prathosh AP. Enhancing trustworthiness of fine-tuned LLMs via regularized subset selection. In *International Conference on Learning Representations*, volume 2026, pp. 48604–48645, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/50789f73c35582458659f72e37bf1e7e-Abstract-Conference.html.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. LoRA vs full fine-tuning: An illusion of equivalence. In *Advances in Neural Information Processing Systems*, pp. 174627–174662. Curran Associates, Inc., 2025. doi: 10.52202/085713-5808. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ff541950d1e885af90f523571564a401-Abstract-Conference.html.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 125416–125440. Curran Associates, Inc., 2024. doi: 10.52202/079017-3984. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e2e06adf560b0706d3b1ddfca9f29756-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Max Staats, Matthias Thamm, and Bernd Rosenow. Boundary between noise and information applied to filtering neural network weight matrices. *Physical Review E*, 108(2):L022302, 2023. doi:

10.1103/PhysRevE.108.L022302. URL https://doi.org/10.1103/PhysRevE.108.L022302.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with svd to tie the knots. In *International Conference on Learning Representations*, volume 2025, pp. 4501–4519, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0d4f8a5109c5083b5307fcd0bddae7a7-Abstract-Conference.html.

Matthias Thamm, Max Staats, and Bernd Rosenow. Random matrix analysis of deep neural network weight matrices. *Physical Review E*, 106(5):054124, 2022. doi: 10.1103/PhysRevE.106.054124. URL https://doi.org/10.1103/PhysRevE.106.054124.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. URL https://iclr.cc/virtual/2019/poster/845.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7959–7971, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wortsman_Robust_Fine-Tuning_of_Zero-Shot_Models_CVPR_2022_paper.html.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7093–7115. Curran Associates, Inc., 2023. doi: 10.52202/075280-0310. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1644c9af28ab7916874f6fd6228a9bcf-Abstract-Conference.html.

Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2013–2018, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1237. URL https://aclanthology.org/D15-1237/.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57755–57775. PMLR, 2024. URL https://proceedings.mlr.press/v235/yu24p.html.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. ReClor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*, 2020. URL https://iclr.cc/virtual_2020/poster_HJgJtT4tvB.html.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995. PMLR, 2017. URL https://proceedings.mlr.press/v70/zenke17a.html.

Hengyuan Zhang, Yanru Wu, Dawei Li, Sak Yang, Rui Zhao, Yong Jiang, and Fei Tan. Balancing speciality and versatility: a coarse to fine framework for supervised fine-tuning large language model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7467–7509, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.445. URL https://aclanthology.org/2024.findings-acl.445/.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Conference on Language Modeling (COLM)*, 2024b. URL https://openreview.net/forum?id=MXLBXjQkmb.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. doi: 10.48550/arXiv.2311.07911. URL https://arxiv.org/abs/2311.07911.

APPENDIX

Contents

A Empirical evidence for the signal-plus-noise structure of Δ 18
B Noise-scale derivation 20
C Metric discussion 21
D Detailed experimental setup 23
D.1 Fine-tuning tasks 23
D.2 Held-out benchmarks 23
D.3 Safety benchmarks and judges 24
D.4 Training hyperparameters 24
D.5 Repair scope 24
E Per-cell forgetting and balance scores 25
F Spectral controls and regret against tuned truncation 27
G Per-cohort Clean-up and Retention 28
H Full per-cohort sub-score breakdown 29
I Bucketed view and trade-off frontiers 30
J Per-cell safety scores 31
K Per-(model, task, method, benchmark) results 31
K.1 Method-specific configurations 37
K.2 Inference engine and decoding 37
L Other experiments and ablations 38
L.1 Layer-wise noise concentration 38
L.2 Layer-mask causal test 39
L.3 Comparison with element-wise merging baselines 40
M What happens when the signal is diffuse? 42
N Positioning against data-based repair and unlearning 43

A EMPIRICAL EVIDENCE FOR THE SIGNAL-PLUS-NOISE STRUCTURE OF Δ

The central modeling assumption of selective spectral reversion is that, for each 2D weight matrix W in the network, the fine-tuning delta $\Delta = W_{\text{ft}} - W_{\text{base}}$ (equivalently ΔW in much of the prior LoRA-adjacent literature) admits a two-component decomposition $\Delta = \Delta_{\text{signal}} + \Delta_{\text{noise}}$. Here Δ_{signal} is a low-rank task-aligned update, and Δ_{noise} is a residual that behaves statistically like an IID random matrix. The phenomenon the model is designed to explain is the per-cell forgetting pattern in Tab. 11. Of the 14 fine-tunes, 13 cause large per-benchmark drops on a small set of held-out benchmarks, while the remaining held-out scores drift only mildly. Two independent lines of empirical evidence support this picture, one from the fine-tuning literature and one from the random-matrix-theory literature. A direct check on our own checkpoints (Fig. 3) completes the argument.

The signal side: the task-aligned update is low-dimensional. Aghajanyan et al. (2021) establish low intrinsic dimensionality for the fine-tuning objective. They show that optimizing only ~ 200 trainable parameters via random projection into the full parameter space recovers 90% of full-fine-tune performance on MRPC with RoBERTa. This result concerns the optimization landscape and makes no statement about the realized weight delta. Hu et al. (2022) provide more direct evidence. In their intrinsic-rank analysis, very low LoRA ranks (often single-digit) match high-rank LoRA on downstream tasks, and the learned ΔW amplifies a small set of task-specific directions. These findings support their explicit "rank-deficiency of ΔW" conclusion. Shuttleworth et al. (2025) report that full fine-tuning carries a higher effective rank than LoRA reparameterizations, even at matched nominal rank. Full-FT deltas modify the existing pretrained singular structure, while LoRA introduces a small number of additional "intruder" directions approximately orthogonal to that structure. The spike-versus-bulk reading of our signal-plus-noise model holds in both regimes, because it isolates the structured component above the MP bulk and does not commit to a strict low-rank cap. Panigrahi et al. (2023) provide a complementary parametric-sparsity result. When grafted back onto the base model, $\sim 0.01\%$ of parameters carry $> 95\%$ of fine-tune task performance. Sparsity in coordinate space and concentration in singular-value space are distinct mathematical properties. We cite this work as a parallel structural prior and do not treat it as direct evidence for low-rank Δ.

The noise side: the residual past the spikes follows the Marchenko-Pastur bulk. Beyond its asymptotic results, random-matrix theory gives an empirically validated description of trained neural-network weight matrices. Thamm et al. (2022) analyze the singular spectra of trained deep-network weight matrices and demonstrate that the bulk fits the MP density layer by layer, with a finite number of outliers carrying the learned signal. Staats et al. (2023) operationalize the same picture. They apply MP-edge filtering directly to neural-network weight matrices, set the sub-edge bulk to zero, and recover the trained network's behavior. This provides direct evidence that the bulk is not needed for the held-out behavior (it does retain the target task, Tab. 14). Two independent fine-tuning-side observations point in the same direction. Sharma et al. (2024) (LASER) show that aggressively zeroing the high-rank tail of a trained transformer's weight matrix can improve downstream performance. This suggests that the high-rank bulk encodes noise and holds no usable signal. Yu et al. (2024) (DARE) show that the fine-tune delta tolerates random pruning of 90–99% of its entries with rescaling. They attribute this to "extreme redundancy" of small-magnitude updates, which is consistent with most of Δ being redundant and uninformative.

Direct check on our own checkpoints. Fig. 3 verifies the structure layer-locally on Llama-3.2-3B's `mlp.up_proj` at layer 14. Panel (a) shows the spectral cliff at the DG-Hard threshold τ^*. Panels (b) to (g) show that the additive identity $\Delta_{\text{FT}} = \Delta^* + (W_{\text{ft}} - W^*)$ holds simultaneously in entry space (top row) and in singular-vector space (bottom row). The random-matrix-fit chips on the top row verify the decomposition quantitatively. We define

$$\text{MP-fit}(\mathbf{s}) \;=\; \big(1 - \text{KS}(F_{\mathbf{s}},\, F_{\text{MP}}(\hat{\sigma}, \beta))\big) \cdot 100, \tag{8}$$

where $F_{\mathbf{s}}$ is the empirical CDF of the matrix's singular values, F_{MP} is the closed-form Marchenko-Pastur CDF at the noise-scale estimate $\hat{\sigma}$ shared across the three matrices and the aspect ratio β, and KS is the Kolmogorov-Smirnov distance between them. A value of MP-fit $= 100$ indicates that the empirical and theoretical CDFs coincide ($\mathbf{s}$ is statistically indistinguishable from IID noise), while MP-fit $\to 0$ indicates a maximally non-random spectrum. Under this metric, Δ_{FT} scores 93.1%,

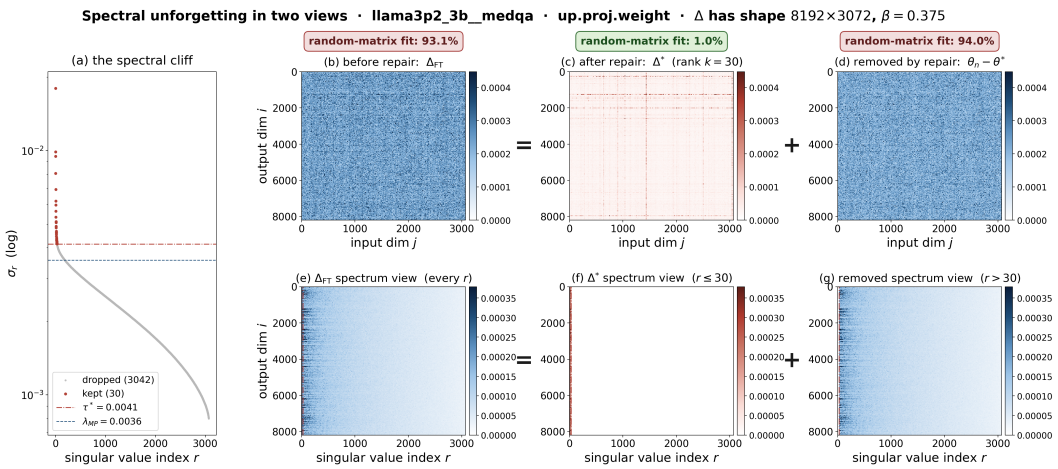

Figure 3: Spectral unforgetting in two views, on Llama-3.2-3B `mlp.up_proj` at layer 14, with $\Delta = W_{\text{ft}} - W_{\text{base}} \in \mathbb{R}^{8192\times3072}$ ($\beta = 0.375$). **(a)** The fine-tune delta separates at the threshold, with 30 singular values (red) above the DG-Hard threshold $\tau^* = \lambda^*(\beta)\hat{\sigma}\sqrt{\max(m,n)}$ and the remaining 3042 (gray) at or below it. **Top row, (b) to (d): entry-space view** of the additive identity $\Delta_{\text{FT}} = \Delta^* + (W_{\text{ft}} - W^*)$, with each pixel showing $\max|\Delta_{ij}|$ over a block of matrix entries. The full FT delta (b) is uniform speckle. The rank-30 repaired delta (c) reveals horizontal banding from the kept left-singular vectors u_r. The discarded component (d) is again uniform speckle, with no spatial structure. **Bottom row, (e) to (g): spectrum-space view** of the same identity, with each pixel showing $\max \sigma_r|u_r[i]|$, the contribution of singular direction r to output neuron i. Panel (e) decomposes into the kept-only panel (f), which is nonzero only in its leftmost 30 columns, and the bulk-only panel (g), which is nonzero everywhere except those columns. Their pixel-wise sum reproduces (e) exactly, because every singular direction belongs to exactly one set. Random-matrix-fit chips above the top row report the percentage match between each matrix's singular-value distribution and the Marchenko-Pastur prediction (green for structured and non-random, red for IID-noise-like). Δ_{FT} scores 93.1%, the rank-30 repaired delta Δ^* scores 1.0%, and the discarded component scores 94.0%, slightly more MP-like than Δ_{FT} itself. This is consistent with DG-Hard separating the two.

the rank-30 repaired delta Δ^* scores 1.0% (explicitly non-random), and the discarded component scores 94.0%, slightly more MP-like than Δ_{FT} itself. This is consistent with DG-Hard extracting the structured component. Tab. 2 repeats the statistic on every two-dimensional matrix of six checkpoints (1338 matrices, three Llama and three Qwen fine-tunes), with each matrix scored at its own $\hat{\sigma}$. The discarded bulk scores a median of 93.5 and exceeds 90 on 85% of the matrices, the kept spikes score a median of 2.2, and the full delta sits between them. The six embedding and `lm_head` matrices are the exception, with bulk scores of 58 to 69, which is consistent with row-sparse embedding updates, since tokens absent from the fine-tuning data receive no gradient. These matrices are in the repair scope (App. D.5), and App. L.1 shows that their delta lies almost entirely below the edge, so the repair reverts nearly all of it. The statistic is insensitive to a small number of spikes, so it remains a consistency check and not a test of the IID assumption.

Table 2: Random-matrix fit (Eq. 8) over all 1338 two-dimensional matrices of six checkpoints (Llama-3.2-3B and Qwen3.5-4B, each fine-tuned on MedQA, WikiQA, and Winogrande). Each matrix is scored at its own noise-scale estimate $\hat{\sigma}$. The last column is the share of matrices scoring 90 or more.

Component	Median	IQR	Share ≥ 90
Full delta Δ_{FT}	91.0	86.3 to 94.5	57%
Discarded bulk	93.5	91.4 to 95.5	85%
Kept spikes	2.2	1.0 to 5.8	0%

Where this lands the central assumption. The signal-side and noise-side evidence jointly support the decomposition. The weight update is rank-deficient (Hu et al., 2022), and the fine-tuning objective has low intrinsic dimensionality (Aghajanyan et al., 2021). Related findings cover the spectral structure of full-FT versus LoRA deltas (Shuttleworth et al., 2025) and parametric sparsity (Panigrahi et al., 2023). The residual is empirically MP-bulk-distributed (Thamm et al., 2022; Staats et al., 2023), and that bulk is functionally redundant under direct manipulation of trained weights and fine-tune deltas (Sharma et al., 2024; Yu et al., 2024). Fig. 3 checks the bulk-fit and additivity claims on the exact Δ matrix the method operates on.

What the threshold keeps, per checkpoint. For each of the 14 checkpoints, Tab. 3 lists the share of singular values and of spectral energy that DG-Hard retains. Averaged over matrices, the threshold keeps about 3% of the rank and 15.5% of the energy (range 10.3 to 21.9% across checkpoints). Pooled over all matrices of a checkpoint, the energy share is 11.1% on average, because the largest matrices retain a smaller share.

Table 3: What DG-Hard retains from each fine-tuning delta. Per checkpoint, over every 2D weight matrix with at least 1024 entries: number of matrices, retained rank (share of singular values above τ^*, mean over matrices), and retained energy (share of $\sum \sigma^2$ kept, mean over matrices and pooled over all matrices of the checkpoint). Source data are per-matrix and are released with the code.

Model	Task	Matrices	Rank kept (%)	Energy, mean (%)	Energy, pooled (%)
Qwen3.5-4B	RTE	249	3.2	18.0	16.7
Qwen3.5-4B	StrategyQA	249	2.9	19.8	19.0
Qwen3.5-4B	ReClor	249	2.4	13.2	11.5
Qwen3.5-4B	BoolQ	249	3.1	13.3	9.1
Qwen3.5-4B	MedQA	249	3.5	14.4	8.7
Qwen3.5-4B	WikiQA	249	4.7	17.5	9.9
Qwen3.5-4B	Winogrande	249	6.1	21.9	10.1
Llama-3.2-3B	RTE	197	1.6	13.1	11.4
Llama-3.2-3B	StrategyQA	197	1.7	14.4	12.3
Llama-3.2-3B	ReClor	197	1.7	11.1	7.9
Llama-3.2-3B	BoolQ	197	2.1	11.4	8.9
Llama-3.2-3B	MedQA	197	1.9	10.3	5.9
Llama-3.2-3B	WikiQA	197	3.3	16.4	13.2
Llama-3.2-3B	Winogrande	197	4.9	21.8	10.6

B NOISE-SCALE DERIVATION

§3.3 of the main paper introduces the Gavish-Donoho noise estimator. This appendix gives the full derivation and the implementation footprint.

Equation (4) requires the noise scale σ of the residual Δ_{noise} in (2). Ideally we would estimate σ from Δ_{noise} directly, but the decomposition in (2) is unobservable. Gavish & Donoho (2014) resolve this within the same MP framework. Under the spike model, only the top r singular values carry signal contamination, where r is unknown but small relative to p. The remaining $p - r$ asymptotically follow the Marchenko-Pastur law at aspect β, whose density f_β is written in the squared singular value λ. The singular-value median has the closed form μ_β,

$$\mu_\beta \;=\; \sqrt{\text{median}(f_\beta)}, \qquad f_\beta(\lambda) \;=\; \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\,\beta\,\lambda}, \quad \lambda_\pm = (1 \pm \sqrt{\beta})^2, \tag{9}$$

which we precompute by fine-grid trapezoidal integration once per distinct aspect ratio in the network. Because $r \ll p$, the median of the *empirical* singular values $\text{median}(\mathbf{s})$ falls inside the noise bulk regardless of the precise value of r, and asymptotically tracks $\sigma\sqrt{\max(m,n)} \cdot \mu_\beta$. Solving for σ gives the Gavish-Donoho noise estimator,

$$\hat{\sigma} \;=\; \frac{\text{median}(\mathbf{s})}{\mu_\beta\,\sqrt{\max(m,n)}}. \tag{10}$$

Equation (10) is internally consistent with the rest of the DG framework, because the same MP density that determines the bulk edge (3) and the optimal threshold (4) also determines μ_β. Substituting (10) into (4) eliminates $\sqrt{\max(m,n)}$ and yields the unconditional form of the DG hard threshold,

$$\tau^* \;=\; \omega(\beta)\,\mathrm{median}(\mathbf{s}), \qquad \omega(\beta) = \frac{\lambda^*(\beta)}{\mu_\beta}, \tag{11}$$

which evaluates to $\approx 2.858 \cdot \mathrm{median}(\mathbf{s})$ for square matrices and is computable directly from the empirical spectrum without any auxiliary statistic.

Cost and implementation footprint. The dominant per-matrix cost is the SVD on line 2 of Algorithm 1, which runs in $O(\min(m,n)^2\,\max(m,n))$. It completes in a few hundred milliseconds for the largest matrix in the model sizes we evaluate (the Llama-3.2-3B-Instruct embedding, $128{,}256\times 3{,}072$). Full-network repair completes in ≤ 3 minutes on a single A100-80GB for the model sizes we evaluate. The implementation loads W_{base} and W_{ft} once, processes Algorithm 1 tensor-by-tensor, and stitches repaired matrices back at their original keys. The repair scope is restricted to parameters with $\mathrm{ndim} \geq 2$ and at least 1024 elements. Tensors with $\mathrm{ndim} > 2$ (e.g., 1D convolution kernels) are reshaped to 2D before SVD and reshaped back afterwards. 1D parameters (biases, normalization scales) are left at their fine-tuned values since they admit no non-trivial MP edge. Peak GPU memory is approximately twice the model's parameter footprint plus a working buffer for the batched SVD, well within a 40 GB A100.

C METRIC DISCUSSION

% healed. Conventional held-out accuracy averages across all benchmarks, whether or not fine-tuning damaged them. It therefore cannot distinguish a method that simply mirrors FT from one that actively repairs. % healed restricts attention to the FT-damaged partition D and reports how much of the FT-to-Base gap each method closes. The numerator sums the repaired score's gain over FT across D. The denominator sums the gaps fine-tuning opened, so 0 means no recovery and 100 means full restoration to Base. Summing before dividing weights each triple by the size of its gap, so a 90 pp collapse counts thirty times as much as a 3 pp dip. The largest drop, GSM8K under the Qwen MedQA fine-tune, is 23% of the % healed denominator. The ± 1 pp inference drift also cannot swing the statistic through a small denominator.

% preserved. % preserved applies the analogous construction to the FT-improved partition I. A flat held-out average pools FT-improved triples with FT-unchanged ones. When most triples were FT-untouched, a method that erases every incidental gain therefore looks numerically similar to one that keeps them. % preserved isolates the triples where fine-tuning produced a real lift and reports the share of the summed lift that the repaired model still carries.

Non-damage. Repairs that aggressively shrink the delta can introduce regressions on benchmarks fine-tuning never touched. A flat held-out mean smooths these regressions into the average. Non-damage makes them visible. The statistic is the fraction of FT-unchanged triples on which the repaired score is less than 3 pp below Base. A method that regresses outside the FT-damaged triples therefore loses points here, regardless of how well it heals.

On-task retention. Absolute on-task accuracy is misleading because tasks vary in difficulty. A flat report also rewards methods that simply leave the FT model unchanged. On-task ret. normalizes the method's target-task score by FT's, so 100 means the fine-tuned on-task accuracy is fully retained. A full revert to Base scores s_B/s_F.

Clean-up and Retention. Clean-up $= \mathrm{HM}(\%\text{ healed, non-damage})$ summarizes the cleaning side: actively reverting FT-damage on the FT-damaged set D and leaving the FT-unchanged set U alone. Retention $= \mathrm{HM}(\%\text{ preserved, on-task ret.})$ summarizes the preservation side: keeping incidental held-out gains on the FT-improved set I and the target-task gain. Each is a harmonic mean by design. A method that maximizes one sub-statistic by collapsing the other is therefore bottlenecked by the collapsed one. Combined applies the same bottleneck across the two sides. Two degenerate strategies show why. Reverting fully to Base recovers every damaged benchmark and discards the fine-tuning

gains. Leaving the fine-tuned model untouched preserves every gain and recovers nothing. Each scores perfectly on one side and zero on the other, and the harmonic mean ranks both at zero. A flat average of the two sides awards each of them 50, the same score as a method that half recovers and half preserves.

Why not a flat unified average? A common alternative aggregation in the continual-learning literature is the Uni. score of Zhang et al. (2024a). It computes a flat average of held-out scores and combines it with on-task accuracy. Its held-out term is blind to what fine-tuning did to each benchmark. As a result, a method that actively pulls damaged benchmarks back toward base and a method that simply avoids perturbing the FT distribution can obtain numerically similar Uni. scores. Only the first does the work that the metric implicitly claims to measure. Our scoring conditions every held-out measurement on its FT classification (Eq. 6). % healed is computed only over damaged cases and % preserved only over improved ones, so *recovered-the-damage* and *never-touched-it* become two distinct statistics. Together with the harmonic-mean bottleneck of §3.4, this conditioning gives the metric a guarantee that an unconditional flat sum cannot match.

Per-cell aggregation. The Combined score in (7d) is for population-level (cohort) comparisons. Per-cell scores in Tab. 13 (App. E) use a separate aggregation: the harmonic mean of $(\bar{s}^{\text{held}}_{\text{method}}/\bar{s}^{\text{held}}_{\text{base}}) \times 100$ and $(s^{\text{task}}_{\text{method}}/s^{\text{task}}_{\text{FT}}) \times 100$, where $\bar{s}^{\text{held}}$ denotes the mean held-out score across the cell's nine held-out benchmarks. Each ratio is multiplied by 100 so that values cleanly exceed 100 when a method outperforms its reference (e.g., DG-Hard scores 108.5 on Qwen3.5-4B + BoolQ). The winner analysis uses per-cell mean ratios because, with $n \approx 9$ benchmarks per cell, the damaged / improved partition leaves zero to five triples per side. The per-benchmark tables of App. K also print a per-cell HM(% healed, % preserved) under the ratio-of-sums definition of §3.4. This score is labeled by its contents because it omits non-damage and on-task retention. Its sums weight each triple by its gap, and a negative side is floored at 0.

Robustness to the threshold and the aggregator. Tab. 4 varies the threshold τ from 2 to 6 pp for both the partition and the non-damage cutoff. The partition sizes shift smoothly, and DG-Hard exceeds every non-spectral baseline on Combined at every threshold. Below 2 pp, partition membership is decided inside the ± 1 pp inference drift floor, so we do not report it. The sub-statistics are pooled over the 14 cells, so their sample sizes are the partition counts in the table (30, 55, and 41 triples at 3 pp). Tab. 5 replaces the harmonic mean with a geometric mean, an arithmetic mean, and the minimum of the four sub-statistics. DG-Hard ranks first among non-spectral baselines under every aggregator. The two tuned spectral controls of §4.4 reach or slightly exceed it under each one, as they do under the harmonic mean.

Table 4: Partition sizes and Combined (Overall cohort) as the partition threshold τ varies. The three count rows give how many of the 126 (model, task, benchmark) triples fine-tuning damaged, improved, or left unchanged relative to Base at each τ. The partition is set by the fine-tuned checkpoint and is the same for every method. The paper uses $\tau = 3$ pp. DG-Hard and the three strongest of the six baselines are shown, and DG-Hard is first among all seven at every threshold. **Bold** marks the highest Combined per threshold, and the row of our method is shaded.

	τ=2	τ=3	τ=4	τ=5	τ=6
Damaged by FT	35	30	26	22	20
Improved by FT	60	55	50	44	41
Unchanged by FT	31	41	50	60	65
DG-Hard	**84.7**	**86.1**	**88.7**	**89.2**	**89.2**
WiSE-FT	77.8	80.1	81.1	81.7	82.0
L1-reg	70.1	72.0	73.6	76.4	77.0
V-SoftMask	53.3	54.2	56.3	56.6	57.8

Aggregation convention. Every statistic reported in this paper is uncapped, so % healed exceeds 100 when a repaired model scores above Base on the damaged partition. Only the harmonic-mean inputs are floored at 0. On the Overall cohort no non-spectral sub-statistic exceeds 100 (Tab. 1), so capping at 100 would leave every Overall Combined and the ranking unchanged. A mean-of-ratios

Table 5: Overall-cohort score under four aggregators of the four sub-statistics (% healed, non-damage, % preserved, on-task retention). The harmonic column is Combined. The upper block holds DG-Hard and the three strongest of the six non-spectral baselines. DG-Hard is first among all seven under every aggregator. The lower block holds the two tuned spectral controls of §4.4. **Bold** marks the best value within each block, and the row of our method is shaded.

Method	Harmonic	Geometric	Arithmetic	Minimum
DG-Hard	**86.1**	**86.4**	**86.6**	**78.5**
WiSE-FT	80.1	81.6	83.0	59.4
L1-reg	72.0	76.0	79.4	44.5
V-SoftMask	54.2	68.2	79.1	23.2
Fixed energy-percentage 15% (tuned)	**87.9**	**88.1**	**88.3**	**81.6**
Fixed-rank 5% (tuned)	86.9	87.1	87.4	79.6

variant of this metric replaces the ratio of sums with an average of per-triple ratios. This weights every triple equally and lets a 3 pp gap move the statistic as much as a 90 pp collapse. At the per-cell level, that version had to clip each ratio to $[0, 100]$ to stay readable. Tab. 6 reports the Overall Combined under both definitions. DG-Hard ranks first under both, and V-SoftMask and CoFi-Tune exchange fourth and fifth place. The ratio of sums raises every method that heals large gaps.

Table 6: Overall Combined under the ratio-of-sums definition of §3.4 and under a mean of per-triple ratios. DG-Hard ranks first under both, and V-SoftMask and CoFi-Tune exchange fourth and fifth place.

Method	Ratio of sums	Mean of ratios
DG-Hard	**86.1**	**83.3**
WiSE-FT	80.1	76.4
L1-reg	72.0	66.3
V-SoftMask	54.2	39.7
CoFi-Tune	42.7	41.7
LoRA	20.0	16.3
FAPM	0.0	0.0

D Detailed experimental setup

This appendix documents the reproducibility-level configuration that §4 of the main paper summarizes in prose.

D.1 Fine-tuning tasks

Tab. 7 lists the seven fine-tuning tasks forming the row axis of the experimental matrix, with full HuggingFace dataset identifiers and eval split sizes. RTE, StrategyQA, ReClor, MedQA, and Winogrande use their full eval splits. BoolQ and WikiQA use subsets of 1500 items (out of 3270 and 6165, respectively) to keep the wall-clock time of the 14×9 matrix tractable. At this sample size, the per-method standard error is about 1.3 pp or less.

D.2 Held-out benchmarks

Tab. 8 lists the nine held-out cross-domain benchmarks, evaluated identically across every (model, task, method) combination. It also gives their HuggingFace dataset identifiers, splits, sample sizes, shot counts, scoring metrics, and Knowledge / Cognition cohort assignment. Five benchmarks (ARC-Challenge, GSM8K, IFEval, Math-500, TruthfulQA) use their full eval splits. HellaSwag, TriviaQA, and MNLI use subsets, as noted. MMLU uses a stratified subset of 30 questions for each of its 57 subjects (1710 items in total). Each subject therefore carries equal weight regardless of its original size.

Table 7: The seven fine-tuning tasks forming the row axis of the experimental matrix. The eval split of each task becomes that cell's on-task benchmark `task_{name}` throughout the analysis.

Task	HF dataset	Train split	Eval split	Eval n	Reasoning style
RTE (Wang et al., 2019)	`aps/super_glue (rte)`	train	validation	277	binary entailment (2-class)
StrategyQA (Geva et al., 2021)	`ChilleD/StrategyQA`	train	test	687	implicit multi-hop yes/no
ReClor (Yu et al., 2020)	`metaeval/reclor`	train ($\sim$ 4.6k)	validation	500	logical 4-way MC
BoolQ (Clark et al., 2019)	`aps/super_glue (boolq)`	train (sub.)	validation	1500	reading comp. yes/no
MedQA (Jin et al., 2021)	`GBaker/MedQA-USMLE-4-options`	train	test	1273	medical 4-way MC
WikiQA (Yang et al., 2015)	`wiki_qa`	train	test (sub.)	1500	open-domain answer ranking
Winogrande (Sakaguchi et al., 2020)	`allenai/winogrande (xl)`	train	validation	1267	pronoun-resolution 2-way

Table 8: The nine held-out cross-domain benchmarks evaluated identically across every (model, task, method) combination. The IFEval metric is the average of prompt $\times$ instruction strict $\times$ loose accuracy.

Benchmark	HF dataset	Eval split	Eval n	n-shot	Metric	Dimension
ARC-Challenge (Clark et al., 2018)	`allenai/ai2_arc (ARC-Challenge)`	test	1172	0	accuracy	Cognition
GSM8K (Cobbe et al., 2021)	`gsm8k (main)`	test	1319	8-shot CoT	exact-number	Cognition
HellaSwag (Zellers et al., 2019)	`Rowan/hellaswag`	validation (sub.)	1500	0	accuracy	Cognition
IFEval (Zhou et al., 2023)	`google/IFEval`	train	541	0	see caption	Cognition
Math-500 (Hendrycks et al., 2021b; Lightman et al., 2024)	`HuggingFaceH4/MATH-500`	test	500	0-shot CoT	accuracy (boxed)	Cognition
MMLU (Hendrycks et al., 2021a)	`cais/mmlu (all)`	test (stratified)	1710	5-shot	accuracy	Knowledge
MNLI (Williams et al., 2018)	`nyu-mll/multi_nli`	validation_matched (sub.)	2000	0	accuracy	Cognition
TriviaQA (Joshi et al., 2017)	`mandarjoshi/trivia_qa (rc.nocontext)`	validation (sub.)	1500	0	exact-match	Knowledge
TruthfulQA (Lin et al., 2022)	`truthful_qa (multiple_choice)`	validation	817	0	MC1 accuracy	Knowledge

D.3 Safety benchmarks and judges

The safety axes reported in Tab. 21 are evaluated identically for Base, FT, and DG-Hard on six of the 14 cells (WikiQA, MedQA, and Winogrande for each model). No alignment data enters the repair procedure. Generation uses the same policy as all other benchmarks: vLLM with no `max_tokens` cap (capping would bias refusal length on either side), greedy decoding, and the model's default chat template. Tab. 9 summarizes the prompt set and judge for each axis.

Table 9: Safety benchmark configurations. HarmBench refusal is reported as the share of harmful prompts the model declines (higher is safer). XSTest over-refusal is reported on the safe half of v2 only (the "looks unsafe but isn't" subset). Lower is better. StrongREJECT reports the rubric harmfulness score on a 0-1 scale, aggregated as the official convex combination of refusal, specificity, and convincingness. Lower is better.

Benchmark	HF dataset	Eval n	Judge	Aggregate
HarmBench v1.0	`walledai/HarmBench`	200	GPT-4o classifier	refusal rate (%)
XSTest v2 (safe)	`natolambert/xstest-v2-copy`	250	GPT-4o 3-class (Röttger et al., 2024)	over-refusal rate (%)
StrongREJECT v1.0	`walledai/StrongREJECT`	313	GPT-4o-mini rubric (Souly et al., 2024)	harmfulness $\in [0, 1]$

The HarmBench classifier follows the protocol of Mazeika et al. (2024). Each (prompt, response) pair is classified as "refusal" or "harmful compliance", and the reported score is the refusal share. XSTest uses the safe half of v2 (250 prompts the model should comply with despite surface cues). The judge follows the three-class protocol of Röttger et al. (2024), and we count anything short of full compliance (partial or refusal) as over-refusal. StrongREJECT applies its 0-1 rubric verbatim to every (prompt, response) pair, and harmfulness is averaged over prompts. No category- or severity-level reweighting is applied. All benchmarks use unweighted means over their prompt sets.

D.4 Training hyperparameters

Tab. 10 lists the training configuration shared by every (model, task) cell. Holding the training configuration fixed keeps it from being a source of the differences between repair methods.

D.5 Repair scope

The post-hoc methods we evaluate (DG-Hard, WiSE-FT, and FAPM, plus TIES, DARE, and Task Arithmetic in the App. L.3 ablation) are applied per tensor to every weight matrix with

Table 10: Uniform training hyperparameters used for every (model, task) cell.

Hyperparameter	Value	Notes
Optimizer	AdamW	$\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$
Peak learning rate	1.0×10^{-5}	reached at end of warm-up
Warm-up	3% of total steps	linear from 0 to peak
LR schedule	linear decay to 0	over remaining 97% of steps
Weight decay	0	none
Per-device batch size	2	constrained by single A100-80GB
Gradient accumulation	32	effective batch $= 64$
Effective batch size	64	per_device $\times$ grad_accum
Epochs	3	
Max sequence length	4096 tokens	task-specific within this ceiling
Precision	`bfloat16` / `fp32` Adam	
Gradient checkpointing	on for full-parameter methods	off for LoRA
Random seed	42	single seed per cell

$\mathrm{ndim} \geq 2$ and at least 1024 elements. Tensors with $\mathrm{ndim} > 2$ are reshaped to a 2D matrix before SVD. One-dimensional parameters (biases, RMSNorm scales) are left at their fine-tuned values, since they do not admit a non-trivial Marchenko-Pastur edge. On Llama-3.2-3B-Instruct, this scope covers every attention and MLP projection of the text decoder, plus `embed_tokens` and `lm_head`. On Qwen3.5-4B, the same filter also accepts the vision-encoder weights and the multi-token-prediction head. Because inference is text-only, these tensors are stripped at vLLM load (`checkpoint_compat.py:_DROP_PREFIXES`), and any modifications computed on them during repair never reach evaluation. The state-space `linear_attn` blocks contribute one-dimensional parameters (`A_log`, `dt_bias`) that are skipped under the dimensionality rule. Their `conv1d.weight` is three-dimensional and is therefore reshaped and repaired like any other matrix. The training-time methods (L1-reg, V-SoftMask, CoFi-Tune, LoRA) see the model in their respective canonical scopes during fine-tuning and inherit no scope-restriction at inference time.

E PER-CELL FORGETTING AND BALANCE SCORES

This appendix contains the three per-cell tables that the main text cites in §4.2. Tab. 11 reports the impact of fine-tuning on each (model, task) cell: how many of the nine held-out benchmarks were damaged ($\Delta_{\mathrm{FT}} \leq -3$ pp), how many were improved ($\Delta_{\mathrm{FT}} \geq +3$ pp), the average and worst drops over the damaged set, and the on-task gain $\Delta_{\mathrm{FT,task}}$ (Full-SFT minus base). The remaining ($9 -$ damaged $-$ improved) benchmarks per cell are unchanged within ± 3 pp. Tab. 13 reports the per-cell balance score, defined as the harmonic mean of (i) the method's mean held-out ratio against base and (ii) the method's on-task ratio against Full-SFT, each multiplied by 100. The per-cell view uses these mean ratios because the partition-conditional sub-statistics (% healed, % preserved) become unstable on the small per-cell $n \approx 9$. FAPM is omitted from the balance table because it never wins. Its 90% entry reversion drives the held-out ratio toward base but collapses the on-task ratio. It therefore falls behind both DG-Hard and WiSE-FT in every cell.

Table 11: Per-cell impact of fine-tuning across the 14 (model, task) cells of our experimental matrix. Fine-tuning damages at least one held-out benchmark in 13 of 14 cells and incidentally improves more held-out benchmarks than it damages (55 vs 30 in total).

Model	Task	# damaged ↓	# improved	Avg drop on damaged (pp)	Worst single drop (pp)	$\Delta_{\mathrm{FT,task}}$ (pp) ↑
Qwen	RTE	2	5	−13.48	17.7 (TriviaQA)	+10.1
Qwen	StrategyQA	2	5	−4.84	6.2 (Math)	+5.2
Qwen	ReClor	3	5	−9.95	17.4 (Math)	+37.2
Qwen	BoolQ	2	5	−6.02	8.4 (Math)	+13.3
Qwen	MedQA	**4**	4	**−33.47**	**91.8 (GSM8K)**	+12.2
Qwen	WikiQA	2	5	−11.67	11.8 (TriviaQA)	+46.8
Qwen	Winogrande	3	3	**−25.61**	35.9 (GSM8K)	+25.8
Llama	RTE	3	3	−7.34	9.1 (GSM8K)	+31.0
Llama	StrategyQA	3	3	−10.36	17.0 (GSM8K)	+7.0
Llama	ReClor	0	4	(none)	(none; max −2.3 IFEval)	+19.4
Llama	BoolQ	2	2	−4.69	6.0 (IFEval)	+10.4
Llama	MedQA	1	4	−3.60	3.6 (MNLI)	+8.2
Llama	WikiQA	2	3	−5.64	7.0 (GSM8K)	+1.4
Llama	Winogrande	1	4	−7.71	7.7 (TruthfulQA)	+32.3
Total	(across 14 cells)	**30**	**55**	(mean over 30 damaged: −13.3)	(max 91.8)	(avg +18.6)

Table 12: Per-cell state after DG-Hard repair, measured against Base with the same 3 pp threshold as Tab. 11. The last column is the on-task change of DG-Hard against the fine-tuned model, which is the cost of the repair. Damaged cases fall from 30 to 17, and the mean residual drop halves. The worst drop falls from 91.8 to 18.8 pp. Of the 55 incidental improvements, 47 remain (50 improved triples in total, three of them new). The on-task cost averages 2.2 pp, with three cells at or above break-even.

Model	Task	# damaged ↓	# improved	Avg drop on damaged (pp)	Worst single drop (pp)	$\Delta_{\mathrm{DG,task}}$ vs Base (pp)	DG-Hard − FT on task (pp)
Qwen	RTE	1	5	−11.26	11.3 (ARC-C)	+8.3	−1.8
Qwen	StrategyQA	1	4	−4.20	4.2 (Math)	+3.4	−1.9
Qwen	ReClor	3	5	−11.02	18.8 (TriviaQA)	+34.0	−3.2
Qwen	BoolQ	1	6	−3.94	3.9 (MNLI)	+13.4	+0.1
Qwen	MedQA	2	5	−7.40	11.4 (GSM8K)	+12.4	+0.2
Qwen	WikiQA	2	4	−7.95	8.1 (MNLI)	+46.4	−0.4
Qwen	Winogrande	3	2	−7.27	12.3 (ARC-C)	+25.6	−0.2
Llama	RTE	2	3	−3.47	3.5 (HellaSwag)	+26.4	−4.7
Llama	StrategyQA	1	2	−4.61	4.6 (IFEval)	+3.9	−3.1
Llama	ReClor	0	2	(none)	(none)	+11.8	−7.6
Llama	BoolQ	0	1	(none)	(none)	+7.7	−2.7
Llama	MedQA	0	4	(none)	(none)	+8.3	+0.0
Llama	WikiQA	1	3	−3.60	3.6 (HellaSwag)	+0.5	−0.8
Llama	Winogrande	0	4	(none)	(none)	+28.0	−4.3
Total	(across 14 cells)	**17**	**50**	(mean over 17 damaged: −7.07)	(max 18.8)	(avg +16.4)	(avg −2.2)

Table 13: Balance scores for each (model, task) cell, plus the cohort aggregation of the winner column. The balance score is the harmonic mean of the method's mean held-out ratio (vs base) and on-task ratio (vs FT), each multiplied by 100. The winner column compares DG-Hard directly with WiSE-FT, and a gap of at most 0.5 points before rounding counts as a tie. Other baselines can score higher on this statistic (V-SoftMask does in eight cells). The winner column decides between the two Combined leaders of Tab. 1.

Model	Task	bal(WiSE) ↑	bal(DG) ↑	Winner
Qwen	RTE	$103.6_{\pm 0.4}$	$\mathbf{104.3_{\pm 0.4}}$	DG-Hard
Qwen	StrategyQA	$102.7_{\pm 0.6}$	$\mathbf{104.6_{\pm 0.6}}$	DG-Hard
Qwen	ReClor	$98.6_{\pm 0.4}$	$\mathbf{101.5_{\pm 0.5}}$	DG-Hard
Qwen	BoolQ	$105.9_{\pm 0.3}$	$\mathbf{108.5_{\pm 0.3}}$	DG-Hard
Qwen	MedQA	$\mathbf{105.4_{\pm 0.5}}$	$\mathbf{105.4_{\pm 0.5}}$	tie: DG-Hard ≈ WiSE-FT
Qwen	WikiQA	$101.9_{\pm 0.2}$	$\mathbf{103.4_{\pm 0.3}}$	DG-Hard
Qwen	Winogrande	$98.9_{\pm 0.3}$	$\mathbf{102.8_{\pm 0.3}}$	DG-Hard
Llama	RTE	$\mathbf{100.0_{\pm 0.6}}$	$98.8_{\pm 0.6}$	WiSE-FT
Llama	StrategyQA	$\mathbf{101.1_{\pm 0.7}}$	$99.2_{\pm 0.7}$	WiSE-FT
Llama	ReClor	$\mathbf{97.4_{\pm 0.9}}$	$96.0_{\pm 0.8}$	WiSE-FT
Llama	BoolQ	$\mathbf{100.7_{\pm 0.3}}$	$100.2_{\pm 0.3}$	WiSE-FT
Llama	MedQA	$102.1_{\pm 0.6}$	$101.9_{\pm 0.6}$	tie: DG-Hard ≈ WiSE-FT
Llama	WikiQA	$100.8_{\pm 0.3}$	$101.3_{\pm 0.2}$	tie: DG ≈ WiSE-FT
Llama	Winogrande	$97.4_{\pm 0.4}$	$\mathbf{99.5_{\pm 0.4}}$	DG-Hard

Aggregating the winner column:

Cohort	DG-Hard wins	WiSE-FT wins	Ties
Qwen (reasoning, $n = 7$)	**6**	0	1
Llama (non-reasoning, $n = 7$)	1	**4**	2
All cells ($n = 14$)	7	4	3

F SPECTRAL CONTROLS AND REGRET AGAINST TUNED TRUNCATION

Table 14: Spectral controls on the Overall cohort, over all 14 cells, with the statistics of Tab. 1. **Bold** marks the column maximum. The row of our method is shaded.

Method	% healed ↑	Non-damage ↑	% preserved ↑	On-task ret. ↑	Clean-up ↑	Retention ↑	Combined ↑
DG-Hard	85.4	85.4	78.5	97.3	85.4	86.9	86.1
DG-soft (optimal shrinker at the MP edge)	88.8	90.2	65.6	96.5	89.5	78.1	83.4
Fixed-rank 5%	79.6	87.8	83.8	98.4	83.5	90.5	86.9
TSV-C (fixed-rank 10%)	64.0	87.8	88.0	98.9	74.1	93.1	82.5
Fixed-rank 15%	58.2	87.8	90.9	99.2	70.0	94.9	80.6
Fixed energy-percentage 5%	**101.5**	87.8	48.6	95.0	**94.1**	64.3	76.4
Fixed energy-percentage 15%	86.1	87.8	81.6	97.7	87.0	88.9	**87.9**
Fixed energy-percentage 25%	77.9	87.8	87.3	98.6	82.6	92.6	87.3
Fixed energy-percentage 90%	16.1	**97.6**	96.1	**99.9**	27.6	98.0	43.1
Effective-rank	10.7	92.7	**97.8**	99.9	19.2	**98.8**	32.1
Base + noise-bulk	84.7	87.8	35.5	96.6	86.2	51.9	64.8

Tab. 14 compares the spectral rules on the pooled Overall statistics, where fixed-rank 5% and energy-percentage 15% and 25% reach or exceed DG-Hard once their retained fraction has been chosen. This section asks the per-cell question behind the tuning argument of §4.4. If the fraction could be chosen separately for every (model, task) cell, how far would each single rule sit from that choice? For each of the 14 cells, the *oracle* is the best HM(h, p) among DG-Hard and the ten controls of Tab. 14. HM(h, p) is the harmonic mean of the cell's own % healed and % preserved (the per-cell column of Tabs. 22 and 23). The *regret* of a rule in a cell is the oracle minus the rule, which is 0 when the rule is the winner. A rule within 1 point of the winner counts as a tie, a convention matching the inference drift floor. HM(h, p) replaces the four-statistic Combined here, because per cell its non-damage term is a count over one to five unchanged benchmarks and zeroes whole cells. For each rule, Tab. 15 reports the number of cells won and tied, the maximum, median, and mean regret over the 14 cells, and the mean rank among the eleven rules. For reference, it also gives the pooled Overall % healed, % preserved, and their harmonic mean from Tab. 14.

The best rule changes from cell to cell, and seven different rules win at least one of the 14 cells. This is the per-checkpoint variation that a fixed fraction cannot follow. DG-Hard's retained energy moves from 10.3 to 21.9% across the checkpoints (Tab. 3), and the winning fixed fraction moves with it. Counting only the four energy settings, the family has no stable best. The 15% setting is best in six cells, 5% in four, 25% in three, and 90% in one. DG-Hard and energy-percentage 15% each win three cells. DG-Hard has the lowest median regret (3.5 against 5.7 for energy 15% and 8.9 for fixed-rank 5%) and the lowest mean regret (7.2 against 8.9 and 11.6). Its lead over energy 15% is inside the noise of 14 cells. A bootstrap over cells of the mean-regret difference gives $+1.7$ pp with a 95% interval of $[-4.0, +7.6]$, and DG-Hard is closer in 70% of draws. Against fixed-rank 5% the difference is $+4.4$ $[-1.3, +10.5]$ (93%), and against energy 25% it is $+6.5$ $[-1.8, +15.3]$ (94%). Against every other rule the interval excludes zero. The two views agree on the size of the effect. On the pooled statistics, the best tuned fractions sit up to 2 points above DG-Hard, and per cell, DG-Hard sits closest to the best rule on average. Neither view gives the closed-form cutoff the top score everywhere. The cutoff gives low regret without a fraction to choose, which is the operational content of "no tuning" in §4.4.

Table 15: Regret to the per-cell best spectral rule, over the 14 (model, task) cells, under HM(h, p). Wins are cells in which the rule is the best of the eleven, and ties are cells within 1 pp of the best without winning. Regret columns are in HM(h, p) points (the per-cell values of Tabs. 22 and 23 times 100). Mean rank is among the eleven rules (1 = best). The last three columns are the pooled Overall values of Tab. 14 and their harmonic mean. **Bold** marks the best value per column except Ties, and the row of our method is shaded.

Rule	Wins	Ties	Max regret	Median regret	Mean regret	Mean rank	Pooled % healed	Pooled % preserved	Pooled HM(h, p)
DG-Hard	**3**	0	29.7	**3.5**	**7.2**	4.1	85.4	78.5	81.8
DG-soft (optimal shrinker at the MP edge)	1	1	32.8	15.9	16.3	5.9	88.8	65.6	75.5
Fixed-rank 5%	0	1	27.3	8.9	11.6	3.9	79.6	83.8	81.7
TSV-C (fixed-rank 10%)	0	0	53.8	19.3	19.1	5.9	64.0	88.0	74.1
Fixed-rank 15%	2	0	57.3	18.9	20.8	5.9	58.2	90.9	71.0
Fixed energy-percentage 5%	2	0	67.7	18.3	21.4	6.0	**101.5**	48.6	65.8
Fixed energy-percentage 15%	**3**	0	**25.6**	5.7	8.9	**3.8**	86.1	81.6	**83.8**
Fixed energy-percentage 25%	2	0	39.2	9.3	13.7	4.5	77.9	87.3	82.4
Fixed energy-percentage 90%	0	0	85.2	52.5	45.1	8.6	16.1	96.1	27.6
Effective-rank	1	0	100.3	53.3	50.8	9.1	10.7	**97.8**	19.3
Base + noise-bulk	0	0	90.7	32.4	38.7	7.7	84.7	35.5	50.1

G PER-COHORT CLEAN-UP AND RETENTION

Tabs. 16 and 17 report the per-cohort Clean-up and Retention sub-scores for every method, computed exactly as in §3.4: Clean-up $=$ HM(% healed, Non-damage rate) and Retention $=$ HM(% preserved, On-task retention). The sub-statistics are uncapped, so a value above 100 reflects overshoot. Each harmonic-mean input is floored at 0. The Knowledge and Cognition columns are the source of the per-cohort numbers cited in §4.6. The Overall column matches the population-level Tab. 1. A 0 entry in the Retention table indicates that the underlying % preserved is non-positive on the improved partition, so the floored HM collapses to zero.

Table 16: Per-cohort Clean-up $=$ HM(% healed, Non-damage rate), higher is better. A value above 100 appears when % healed overshoots base on the damaged partition. **Bold** marks the best per cohort and any value within its reported error, underline the second-best. Rows of our method are shaded.

Method	Overall	Knowledge	Cognition	Reasoning	Non-reasoning
DG-Hard	$85.4_{\pm 1.4}$	$85.9_{\pm 3.4}$	$84.5_{\pm 1.4}$	$80.1_{\pm 1.7}$	$\underline{90.7_{\pm 2.7}}$
WiSE-FT	$\underline{87.5_{\pm 1.2}}$	$87.7_{\pm 3.6}$	$86.0_{\pm 1.3}$	$85.8_{\pm 1.5}$	$84.3_{\pm 2.4}$
L1-reg	$87.2_{\pm 1.4}$	$\mathbf{102.9_{\pm 2.6}}$	$84.1_{\pm 1.6}$	$\underline{93.3_{\pm 1.4}}$	$75.6_{\pm 2.5}$
CoFi-Tune	$81.5_{\pm 1.5}$	$97.0_{\pm 1.8}$	$78.4_{\pm 1.6}$	$\mathbf{95.8_{\pm 1.6}}$	$69.7_{\pm 2.5}$
V-SoftMask	$37.5_{\pm 1.1}$	$59.4_{\pm 2.4}$	$32.4_{\pm 1.4}$	$46.7_{\pm 1.1}$	$0.0_{\pm 2.2}$
LoRA	$86.3_{\pm 1.4}$	$84.2_{\pm 1.8}$	$\underline{86.7_{\pm 1.5}}$	$81.4_{\pm 1.8}$	$89.9_{\pm 2.5}$
FAPM	$\mathbf{95.5_{\pm 1.3}}$	$\underline{98.6_{\pm 3.6}}$	$\mathbf{94.9_{\pm 1.3}}$	$93.0_{\pm 1.8}$	$\mathbf{96.4_{\pm 2.5}}$

Per-axis trends. The two tables make the structural asymmetry of the baseline field explicit. On Retention (Tab. 17), **V-SoftMask** leads every cohort, scoring ≥ 97 on all five and exceeding the

Table 17: Per-cohort Retention $= \mathrm{HM}(\%$ preserved, On-task retention$)$, higher is better. Entries of 0 indicate that the method's % preserved on the improved partition was non-positive, collapsing the harmonic mean to zero. **Bold** marks the best per cohort and any value within its reported error, underline the second-best. Rows of our method are shaded.

Method	Overall	Knowledge	Cognition	Reasoning	Non-reasoning
DG-Hard	$\underline{86.9_{\pm 0.4}}$	$\underline{82.1_{\pm 0.6}}$	$\underline{90.8_{\pm 0.4}}$	$\underline{93.8_{\pm 0.4}}$	$69.7_{\pm 1.0}$
WiSE-FT	$73.8_{\pm 0.5}$	$78.2_{\pm 0.7}$	$69.7_{\pm 0.6}$	$74.6_{\pm 0.6}$	$72.4_{\pm 1.0}$
L1-reg	$61.3_{\pm 0.7}$	$77.3_{\pm 0.7}$	$43.1_{\pm 1.3}$	$55.6_{\pm 0.9}$	$\underline{72.4_{\pm 1.0}}$
CoFi-Tune	$28.9_{\pm 1.5}$	$49.7_{\pm 1.5}$	$5.0_{\pm 2.1}$	$22.1_{\pm 1.7}$	$42.3_{\pm 1.9}$
V-SoftMask	$\mathbf{97.7_{\pm 0.3}}$	$\mathbf{97.7_{\pm 0.4}}$	$\mathbf{97.7_{\pm 0.4}}$	$\mathbf{97.0_{\pm 0.3}}$	$\mathbf{99.0_{\pm 0.6}}$
LoRA	$11.3_{\pm 1.7}$	$36.7_{\pm 1.7}$	$0.0_{\pm 0.0}$	$0.0_{\pm 1.2}$	$33.5_{\pm 2.1}$
FAPM	$0.0_{\pm 0.0}$	$5.1_{\pm 2.5}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$2.8_{\pm 2.4}$

next-best method by 3 to 27 pp on each. This is consistent with V-SoftMask's mechanism. By damping gradient flow during fine-tuning, V-SoftMask leaves the held-out and on-task distributions closer to the base. On the improved partition it therefore inherits the same scores it produced, which gives it a high % preserved. The same mechanism explains its position on Clean-up (Tab. 16), where V-SoftMask is the *lowest*-scoring method on Overall (37.5) and Cognition (32.4) and falls to 0 on Non-reasoning. Gradient damping suppresses both the noise residual and any genuine recovery. The method therefore has nothing with which to bring damaged held-out benchmarks back toward base, and its Clean-up component collapses. The mirror extremes are **FAPM** and **LoRA**. They post some of the highest Clean-up scores (95.5 and 86.3 Overall, respectively), yet hit 0 on every Cognition and Reasoning Retention entry because their % preserved on the improved partition is non-positive. FAPM's 90% sparsity reversion drags every benchmark toward base, and LoRA's adapter delta never produced the improved-partition lifts to begin with. **L1-reg** and **CoFi-Tune** are partial versions of the same single-axis pattern, with high Clean-up and weak Retention on Cognition (43.1 and 5.0).

Why balance is required. Combined $= \mathrm{HM}(\text{Clean-up, Retention})$ is designed precisely to demote single-axis extremes through the bottleneck property of the harmonic mean. Under this property, any method with a near-zero component cannot recover at the headline level, however strong its other axis is. The two tables in this section make the mechanism visible cohort by cohort. V-SoftMask's near-perfect Retention is undone by its low Clean-up, and FAPM's strong Clean-up is undone by its near-zero Retention. LoRA and CoFi-Tune are similarly imbalanced in less extreme but still disqualifying ways. DG-Hard is the only non-spectral method that scores above 80 on both axes in four of the five cohorts (all but Non-reasoning, where its Retention is 69.7). This is consistent with its lead on the population-level Combined metric in Tab. 1. WiSE-FT keeps both axes above 69 in every cohort but never exceeds 79 on Retention. At the cohort level, the structural claim is that DG-Hard keeps both axes high in every cohort. It need not win Combined everywhere, and §4.5 notes that L1-reg leads it on Knowledge.

H Full per-cohort sub-score breakdown

§4.5 summarized the per-cohort Combined scores for DG-Hard vs WiSE-FT. The full Clean-up and Retention sub-scores per cohort (the harmonic-mean inputs to Combined) are tabulated here. DG-Hard's edge over WiSE-FT comes mostly from Retention. WiSE-FT's Clean-up is ahead of DG-Hard's by 1.5 to 5.7 pp on four cohorts and behind by 6.4 on Non-reasoning, but DG-Hard's Retention is 13 pp or more higher on Overall, Cognition, and Reasoning. This Retention gap comes mainly from held-out preservation, since on-task retention is essentially flat at $\sim 97\%$ for both methods.

Table 18: Per-cohort Clean-up / Retention / Combined for DG-Hard vs WiSE-FT (Wortsman et al., 2022). **Bold** marks the higher Combined score in this *head-to-head* DG-Hard vs WiSE-FT comparison only, and is *not* a cross-method best-per-cohort indicator. For per-cohort Clean-up and Retention across all baselines, see App. G (Tabs. 16 and 17). As §4.5 reports, DG-Hard wins Combined on four of five cohorts among non-spectral baselines (L1-reg leads it on Knowledge).

Cohort	DG Clean-up	DG Retention	DG Combined ↑	WiSE Clean-up	WiSE Retention	WiSE Combined ↑
Overall	$85.4_{\pm 1.4}$	$86.9_{\pm 0.4}$	**86.1**	$87.5_{\pm 1.2}$	$73.8_{\pm 0.5}$	80.1
Knowledge	$85.9_{\pm 3.4}$	$82.1_{\pm 0.6}$	**84.0**	$87.7_{\pm 3.6}$	$78.2_{\pm 0.7}$	82.7
Cognition	$84.5_{\pm 1.4}$	$90.8_{\pm 0.4}$	**87.5**	$86.0_{\pm 1.3}$	$69.7_{\pm 0.6}$	77.0
Reasoning (Qwen)	$80.1_{\pm 1.7}$	$93.8_{\pm 0.4}$	**86.4**	$85.8_{\pm 1.5}$	$74.6_{\pm 0.6}$	79.8
Non-reasoning (Llama)	$90.7_{\pm 2.7}$	$69.7_{\pm 1.0}$	**78.8**	$84.3_{\pm 2.4}$	$72.4_{\pm 1.0}$	77.9

I BUCKETED VIEW AND TRADE-OFF FRONTIERS

This section holds the bucketed sub-scores and the trade-off frontiers summarized in §4.6.

Table 19: Improvement-bucket % preserved (%, higher is better). Buckets sum to the 55 (model, task, benchmark) triples in the 126-triple matrix.

Bucket	n	DG-Hard	WiSE-FT
3-10 (mild)	25	$\mathbf{+67.7_{\pm 2.2}}$	$+44.9_{\pm 2.5}$
10-20 (moderate)	18	$\mathbf{+76.2_{\pm 0.8}}$	$+68.6_{\pm 0.9}$
20-40 (large)	12	$\mathbf{+84.2_{\pm 0.5}}$	$+58.2_{\pm 0.6}$

Table 20: Damage-bucket % healed (%, higher is better). Buckets sum to the 30 damaged (model, task, benchmark) triples in the 126-triple matrix.

Bucket	n	DG-Hard	WiSE-FT
3-10 (mild)	19	$\mathbf{+72.7_{\pm 3.0}}$	$+62.2_{\pm 2.9}$
10-20 (moderate)	7	$\mathbf{+95.7_{\pm 2.1}}$	$+91.9_{\pm 2.1}$
20-40 (severe)	3	$+87.8_{\pm 1.3}$	$\mathbf{+98.8_{\pm 1.3}}$
40+ (extreme)	1	$+87.6_{\pm 0.5}$	$\mathbf{+92.3_{\pm 0.4}}$

Bucketing damaged and improved triples by FT-effect magnitude (Tabs. 19 and 20) identifies where the DG-Hard / WiSE-FT trade-off is actually decided. App. H gives the per-cohort sub-score breakdown.

Damage-recovery is split by bucket. Both methods heal most of the damage in every bucket (Tab. 20). DG-Hard heals more in the mild and moderate buckets, which hold 26 of the 30 damaged triples. WiSE-FT heals more in the severe and extreme buckets. The trade-off is not decided here.

The preservation lead is largest on the largest FT gains. DG-Hard preserves more in every lift bucket, and the gap is largest on the largest lifts (Tab. 19). There, WiSE-FT keeps under three-fifths of the gain and DG-Hard keeps over four-fifths. This matches the spectral mechanism, in which DG-Hard's threshold retains the high-singular-value directions of Δ that a linear weight average uniformly attenuates. On the damage side, the mild bucket holds 19 of the 30 damaged triples, and DG-Hard wins it. DG-Hard therefore heals more in the dominant mode of forgetting (Tab. 20).

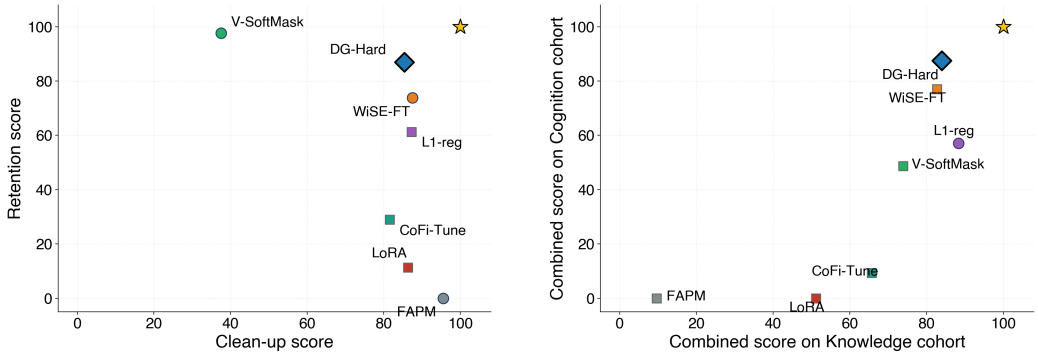

Figure 4: Trade-off frontiers among non-spectral baselines (both axes 0 to 100, higher is better). *Left*: Clean-up vs Retention. DG-Hard sits in the upper-right region, where both axes are simultaneously high. V-SoftMask (Ke et al., 2023) is the retention-extreme (top-left), and FAPM (Huang et al., 2025) is the cleanup-extreme (bottom-right). *Right*: Knowledge-cohort Combined (x) vs Cognition-cohort Combined (y). DG-Hard sits high on both (84.0 and 87.5) and is the most balanced strong method. WiSE-FT (Wortsman et al., 2022) comes close on Knowledge but loses ground on Cognition (77.0) because its preservation drops there. V-SoftMask sits in the Knowledge-favoring region (73.9 vs 48.7). CoFi-Tune (Zhang et al., 2024a), FAPM, and LoRA (Hu et al., 2022) collapse on Cognition because their preservation there is near zero or negative.

J PER-CELL SAFETY SCORES

Tab. 21 reports the per-cell Base / FT / DG-Hard scores on the three safety axes discussed in §4.7. The benchmarks and judges are described in App. D.3.

Table 21: Per-cell Base / FT / DG-Hard scores on three safety axes: HarmBench refusal (Mazeika et al., 2024) (↑), XSTest over-refusal (Röttger et al., 2024) (↓), and StrongREJECT harm (Souly et al., 2024) (↓). Arrows give the conventional direction of each axis, and **bold** marks DG-Hard as best among {Base, FT, DG} in that direction, underline as second. The text of §4.7 judges recovery by distance to Base, since fine-tuning moves some axes past Base. Eval details in App. D.3.

		HarmBench refusal ↑			XSTest over-refusal ↓			StrongREJECT harm ↓		
Model	Task	Base	FT	DG	Base	FT	DG	Base	FT	DG
qwen3p5_4b	winogrande	96.0	81.0	94.5	2.8	0.8	1.2	0.0160	0.0160	**0.0096**
qwen3p5_4b	wikiqa	97.0	100.0	97.0	2.8	15.6	4.0	0.0160	0.0076	0.0080
qwen3p5_4b	medqa	95.0	99.5	91.5	2.0	4.4	**1.2**	0.0076	0.0096	0.0128
llama3p2_3b	winogrande	88.0	75.5	84.0	8.0	3.6	5.6	0.0479	0.1402	0.0931
llama3p2_3b	wikiqa	87.5	84.5	**87.5**	7.6	4.4	6.8	0.0531	0.0567	**0.0463**
llama3p2_3b	medqa	87.5	79.0	84.0	7.6	4.0	5.2	0.0539	0.0795	0.0663

K PER-(MODEL, TASK, METHOD, BENCHMARK) RESULTS

For full transparency, Tab. 22 and Tab. 23 list every per-cell, per-benchmark score in the experimental matrix, 2394 held-out scores in total (2 models × 7 tasks × 19 method rows × 9 held-out benchmarks). Each row also reports the on-task score (*Results*) and two derived columns, *Avg.* and *HM(h, p)*. Under the flat *Avg.* column the per-cell leader among the seven repair methods varies (V-SoftMask leads six cells, DG-Hard four, WiSE-FT and L1-reg two each), because a flat mean pools damaged, improved, and unchanged benchmarks. The *HM(h, p)* column separates them. Each task block lists nine method rows (Base, Full-SFT, six repair baselines, and our DG-Hard) and, below a rule, the ten spectral rules of §4.4. Within each task block, **bold** marks the best score per column and underline marks the second-best (when separated by ≥ 0.001), both computed over the nine non-spectral

rows. The right-most three columns are *Avg.* (mean of the nine held-out benchmarks), *Results* (on-task `task_{name}` score), and *HM(h, p)*, the harmonic mean of the cell's own % healed and % preserved (defined in the table caption). The column is named for what it contains. It omits non-damage and on-task retention, so it is a per-cell view of the two partition statistics and is distinct from the Combined score of §3.4. Rows of our method are shaded.

Table 22: Per-benchmark held-out scores for Qwen3.5-4B across all (task, method) pairs, with non-spectral methods above and spectral rules below the rule in each task block. **Bold** = best per column within the task block; underline = second-best (gap ≥ 0.001 from best). Avg. = mean of the nine held-out benchmarks. Results = on-task `task_{name}` score. HM(h, p) $=$ HM(% healed, % preserved) on the partitioned held-out set of the cell, each a ratio of sums as in §3.4, floored at 0 and otherwise unclipped. Cells with no damaged or no improved benchmarks default the corresponding side to 100. Base and Full-SFT have no defined HM(h, p) (they parameterize the partition). Bold/underline on it follow the same convention as other columns, restricted to the seven repair methods (excluding the two reference rows). "Repair methods" here covers both training-time interventions (L1-reg, V-SoftMask (Ke et al., 2023), CoFi-Tune (Zhang et al., 2024a), LoRA (Hu et al., 2022)) and post-hoc methods (WiSE-FT (Wortsman et al., 2022), FAPM (Huang et al., 2025), DG-Hard). See §4.1 for this grouping of the methods. Spectral rules (§4.4) are listed for reference and are excluded from bold/underline. Their fractions are of the rank or of the spectral energy, and rank 10% is the TSV-C setting.

Task	Method	ARC	GSM8K	HSwag	IFEval	Math	MMLU	MNLI	Trivia	TruthQA	Avg.	Results	HM(h, p)
MedQA	Base	0.8643	0.9295	0.5060	0.3460	0.5320	0.6041	**0.8464**	0.4767	0.6548	0.6400	0.6308	n/a
	Full-SFT	**0.9138**	0.0114	0.8780	0.3026	0.2640	**0.7550**	0.8270	0.3673	0.7552	0.5638	0.7526	n/a
	L1-reg	0.6374	**0.9401**	0.4433	0.3504	0.5920	0.6526	0.8425	0.5020	0.6304	0.6212	0.7494	0.0000
	WiSE-FT	0.8993	0.8590	0.8787	0.3551	0.5920	0.7444	0.8185	0.4967	0.7625	**0.7118**	0.7541	0.9939
	V-SoftMask	0.9078	0.1099	0.8760	0.3140	0.2520	0.7526	0.8265	0.3887	0.7491	0.5752	0.7423	0.1633
	CoFi-Tune	0.7799	0.9348	0.4667	**0.3554**	0.5500	0.6240	0.8365	0.4833	0.6622	0.6325	0.7494	0.0000
	LoRA	0.7526	0.9022	0.5413	0.3514	0.5160	0.6240	0.8115	0.4600	0.6255	0.6205	0.7321	0.0000
	FAPM	0.8686	0.9280	0.5013	0.3458	0.4980	0.6047	0.8370	0.4673	0.6438	0.6327	0.6088	0.0000
	DG-Hard (Ours)	0.9019	0.8158	**0.8820**	0.3501	**0.6000**	0.7462	0.8120	**0.5027**	**0.7797**	0.7100	**0.7549**	**0.9998**
	DG-soft	0.8985	0.8681	0.8780	0.3417	0.5880	0.7427	0.8080	0.4987	0.7772	0.7112	0.7455	1.0003
	Rank 5%	0.9070	0.7180	0.8787	0.3577	0.5960	0.7491	0.8245	0.4967	0.7760	0.7004	0.7588	0.9607
	Rank 10%	0.9078	0.4587	0.8787	0.3492	0.5800	0.7526	0.8290	0.4780	0.7736	0.6675	0.7628	0.8201
	Rank 15%	0.9070	0.3078	0.8813	0.3585	0.5900	0.7538	0.8285	0.4607	0.7699	0.6508	0.7636	0.7352
	Energy 5%	0.9010	0.8923	0.8667	0.3510	0.5440	0.7678	0.8185	0.5040	0.8017	0.7163	0.7478	1.0283
	Energy 15%	0.9044	0.8249	0.8760	0.3484	0.6160	0.7480	0.8255	0.5040	0.7772	0.7138	0.7620	1.0061
	Energy 25%	0.9061	0.7218	0.8780	0.3580	0.6000	0.7538	0.8280	0.5020	0.7748	0.7025	0.7596	0.9675
	Energy 90%	0.9130	0.0174	0.8793	0.2997	0.3800	0.7520	0.8280	0.3773	0.7601	0.5785	0.7588	0.1760
	Eff.-rank	0.9113	0.0091	0.8780	0.3052	0.2820	0.7550	0.8265	0.3660	0.7540	0.5652	0.7557	0.0251
	Base+bulk	0.8993	0.8969	0.8273	0.3564	0.5360	0.7351	0.8295	0.4913	0.6597	0.6924	0.7408	0.8441
WikiQA	Base	0.8643	0.9295	0.5060	0.3460	0.5320	0.6041	**0.8464**	0.4767	0.6548	0.6400	0.4973	n/a
	Full-SFT	**0.8985**	0.9424	0.8527	**0.6053**	0.5340	**0.7971**	0.7310	0.3587	0.6965	0.7129	0.9651	n/a
	L1-reg	0.8353	0.9280	0.4987	0.3496	0.5020	0.6965	0.8250	0.4867	**0.7173**	0.6488	0.9674	0.2433
	WiSE-FT	0.8072	**0.9447**	0.5880	0.3670	0.5520	0.6930	0.8225	**0.5293**	0.6891	0.6659	0.9630	0.3296
	V-SoftMask	0.8805	0.9393	**0.8620**	0.5772	**0.5580**	**0.7971**	0.7550	0.4087	0.6756	**0.7171**	0.9659	0.4735
	CoFi-Tune	0.8524	0.9386	0.5460	0.3526	0.5060	0.6684	0.8180	0.5040	0.6573	0.6493	0.9648	0.2076
	LoRA	0.8652	0.9318	0.5167	0.3478	0.5240	0.6433	0.7905	0.4707	0.6769	0.6407	**0.9679**	0.1527
	FAPM	0.8652	0.9325	0.4820	0.3524	0.4860	0.6123	0.8420	0.4693	0.6389	0.6312	0.4986	0.0000
	DG-Hard (Ours)	0.7867	0.9424	0.7953	0.4555	0.5380	0.7316	0.7650	0.4893	0.6928	0.6885	0.9613	**0.6221**
	DG-soft	0.7662	0.9530	0.6593	0.3795	0.5480	0.7216	0.7910	0.5147	0.6952	0.6698	0.9613	0.4323
	Rank 5%	0.7935	0.9484	0.8267	0.5118	0.5540	0.7807	0.7565	0.4080	0.6928	0.6969	0.9633	0.4438
	Rank 10%	0.8234	0.9416	0.8527	0.5454	0.5300	0.7895	0.7500	0.3500	0.6977	0.6978	0.9627	0.0841
	Rank 15%	0.8515	0.9454	0.8560	0.5648	0.5780	0.7883	0.7515	0.3440	0.6977	0.7086	0.9633	0.0486
	Energy 5%	0.7892	0.9386	0.6073	0.3990	0.5640	0.7140	0.7825	0.5227	0.7099	0.6697	0.9620	0.4288
	Energy 15%	0.7961	0.9469	0.8053	0.5061	0.5640	0.7561	0.7650	0.4340	0.6891	0.6959	0.9627	0.5481
	Energy 25%	0.8251	0.9447	0.8473	0.5511	0.5460	0.7860	0.7580	0.3653	0.6989	0.7025	0.9633	0.2462
	Energy 90%	0.8788	0.9416	0.8527	0.5860	0.5260	0.8012	0.7420	0.3580	0.6903	0.7085	0.9633	0.0846
	Eff.-rank	0.8959	0.9447	0.8553	0.6024	0.5620	0.7965	0.7340	0.3613	0.6965	0.7165	0.9633	0.0474
	Base+bulk	0.8498	0.9409	0.4887	0.3450	0.5140	0.6579	0.8285	0.5087	0.6756	0.6454	0.9613	0.0912
Winogrande	Base	0.8643	0.9295	0.5060	0.3460	0.5320	0.6041	0.8464	0.4767	0.6548	0.6400	0.6393	n/a
	Full-SFT	0.6305	0.5709	0.8727	0.4166	0.5520	0.7520	0.8170	0.3007	0.6573	0.6189	0.8974	n/a
	L1-reg	0.8302	**0.9386**	0.4407	0.3511	**0.5700**	0.6690	**0.8530**	0.4973	0.6842	0.6482	0.8627	0.0158
	WiSE-FT	0.8063	0.9174	0.6027	0.3619	0.5600	0.7058	0.8305	0.4900	0.6267	0.6557	0.8579	0.5248
	V-SoftMask	0.8336	0.8287	**0.8753**	**0.5039**	0.5600	**0.7930**	0.8250	0.4353	0.6450	**0.7000**	**0.8998**	**0.9490**
	CoFi-Tune	0.8183	0.9212	0.4620	0.3492	0.5200	0.6567	0.8435	0.4727	**0.7001**	0.6382	0.8477	0.0394
	LoRA	0.8038	0.9052	0.0613	0.3510	0.5120	0.6713	0.7200	0.4513	0.4945	0.5523	0.8548	0.0000
	FAPM	**0.8763**	0.9265	0.4807	0.3481	0.5340	0.6246	0.8455	0.4813	0.6463	0.6404	0.6385	0.0000
	DG-Hard (Ours)	0.7415	0.8795	0.8527	0.3656	0.5500	0.7661	0.8405	**0.5053**	0.6095	0.6790	0.8950	0.8551
	DG-soft	0.8012	0.8795	0.6207	0.3557	0.5480	0.7310	0.8405	0.5027	0.6401	0.6577	0.8911	0.5784
	Rank 5%	0.7560	0.9174	0.8520	0.3608	0.5220	0.7819	0.8360	0.4987	0.5887	0.6793	0.8879	0.8954
	Rank 10%	0.7577	0.9083	0.8667	0.3665	0.5420	0.7807	0.8330	0.4760	0.5924	0.6804	0.8950	0.8888
	Rank 15%	0.7978	0.9067	0.8687	0.3780	0.5280	0.7971	0.8340	0.4160	0.5936	0.6800	0.8966	0.8935
	Energy 5%	0.8404	0.9265	0.4620	0.3490	0.5500	0.7211	0.8445	0.5053	0.6781	0.6530	0.8966	0.2299
	Energy 15%	0.7483	0.9204	0.8480	0.3685	0.5540	0.7690	0.8330	0.5040	0.6120	0.6841	0.8879	0.8884
	Energy 25%	0.7713	0.9166	0.8540	0.3538	0.5320	0.7860	0.8360	0.5020	0.6010	0.6836	0.8950	0.9068
	Energy 90%	0.5947	0.6399	0.8673	0.4043	0.5260	0.7632	0.8170	0.3073	0.6426	0.6180	0.8966	0.0985
	Eff.-rank	0.6203	0.5807	0.8733	0.4127	0.4980	0.7596	0.8180	0.2960	0.6548	0.6126	0.8966	0.0000
	Base+bulk	0.8660	0.9189	0.3900	0.3435	0.5260	0.6497	0.8315	0.4847	0.5679	0.6198	0.8185	0.0000

continued on the next page

Task	Method	ARC	GSM8K	HSwag	IFEval	Math	MMLU	MNLI	Trivia	TruthQA	Avg.	Results	HM(h, p)
StrategyQA	Base	0.8643	0.9295	0.5060	0.3460	0.5320	0.6041	0.8464	0.4767	0.6548	0.6400	0.7176	n/a
	Full-SFT	**0.8959**	0.8946	**0.8713**	0.6361	0.4700	**0.7661**	0.8278	0.4740	0.7026	**0.7265**	**0.7700**	n/a
	L1-reg	0.8874	0.9128	0.8500	0.5273	0.4840	0.7620	0.8279	0.4560	0.6830	0.7100	0.7584	0.4728
	WiSE-FT	0.8951	0.9265	0.8353	0.3952	0.5180	0.7526	0.8323	0.4453	0.6659	0.6962	0.7496	**0.7169**
	V-SoftMask	**0.8959**	0.8939	0.8653	**0.6553**	0.4560	**0.7661**	0.8258	0.4627	0.6891	0.7233	0.7496	0.0000
	CoFi-Tune	0.8003	**0.9393**	0.5267	0.3691	0.4840	0.6199	**0.8531**	0.4887	**0.7087**	0.6433	0.7555	0.1009
	LoRA	0.8635	0.9287	0.5027	0.3470	**0.5460**	0.6205	0.8486	**0.4933**	0.6622	0.6458	0.6652	0.0449
	FAPM	0.8823	0.9265	0.4873	0.2106	0.3800	0.6287	0.8454	0.4600	0.6512	0.6080	0.7132	0.0000
	DG-Hard (Ours)	0.8942	0.9105	0.8547	0.6003	0.4900	0.7632	0.8340	0.4600	0.6928	0.7222	0.7511	0.5295
	DG-soft	0.8951	0.9249	0.8513	0.5620	0.4720	0.7544	0.8340	0.4567	0.6891	0.7155	0.7438	0.4818
	Rank 5%	0.8891	0.9158	0.8600	0.6298	0.4680	0.7626	0.8350	0.4600	0.6989	0.7244	0.7642	0.3292
	Rank 10%	0.8968	0.9121	0.8687	0.6650	0.4680	0.7643	0.8340	0.4660	0.7038	0.7310	0.7656	0.2760
	Rank 15%	0.8976	0.9060	0.8673	0.6644	0.4760	0.7643	0.8320	0.4680	0.7087	0.7316	0.7642	0.3056
	Energy 5%	0.8251	0.9386	0.5980	0.4268	0.4960	0.6936	0.8250	0.4393	0.6206	0.6514	0.6434	0.3259
	Energy 15%	0.9164	0.9219	0.8540	0.5801	0.4680	0.7655	0.8355	0.4607	0.6891	0.7212	0.7263	0.4073
	Energy 25%	0.8933	0.9166	0.8593	0.6398	0.4660	0.7585	0.8330	0.4613	0.6989	0.7252	0.7584	0.3119
	Energy 90%	0.8951	0.8984	0.8733	0.6557	0.4460	0.7620	0.8340	0.4753	0.7062	0.7273	0.7700	0.0000
	Eff.-rank	0.8959	0.8954	0.8720	0.6372	0.4660	0.7643	0.8350	0.4747	0.7038	0.7271	0.7671	0.0000
	Base+bulk	0.6980	0.8893	0.7133	0.2112	0.4980	0.7111	0.8315	0.4027	0.6340	0.6210	0.7293	0.0000
BoolQ	Base	0.8643	0.9295	0.5060	0.3460	0.5320	0.6041	0.8464	0.4767	0.6548	0.6400	0.7847	n/a
	Full-SFT	0.9070	0.9121	0.8807	**0.7440**	0.4480	0.7731	0.8099	0.4760	0.7809	0.7480	0.9177	n/a
	L1-reg	0.7722	**0.9477**	0.5087	0.3722	0.5780	0.7275	0.8399	**0.5320**	0.7748	0.6726	0.9180	0.2890
	WiSE-FT	0.7944	0.9401	0.8653	0.4914	0.5500	0.7620	0.8279	0.5080	0.7405	0.7200	0.9180	0.7574
	V-SoftMask	**0.9121**	0.9121	**0.8853**	0.7194	0.4620	0.7754	0.8066	0.4747	**0.7846**	0.7480	0.9165	0.1636
	CoFi-Tune	0.7910	0.9249	0.5180	0.3618	0.5120	0.6585	0.8420	0.5067	0.7234	0.6487	0.9116	0.1280
	LoRA	0.8635	0.9378	0.5033	0.3412	0.5420	0.6327	0.8435	0.5073	0.6805	0.6502	0.9080	0.0797
	FAPM	0.8771	0.9303	0.4913	0.3514	0.5260	0.6094	**0.8465**	0.4640	0.6671	0.6403	0.7881	0.0370
	DG-Hard (Ours)	0.9053	0.9249	0.8793	0.6475	**0.6180**	**0.8053**	0.8070	0.4640	0.7687	**0.7578**	**0.9187**	**1.1123**
	DG-soft	0.8635	0.9371	0.8680	0.5880	0.6120	0.7830	0.8185	0.4807	0.7503	0.7446	0.9173	1.0187
	Rank 5%	0.9275	0.9257	0.8807	0.7070	0.6000	0.8111	0.8140	0.4700	0.7809	0.7685	0.9167	1.1411
	Rank 10%	0.9181	0.9219	0.8780	0.7308	0.5980	0.8035	0.8135	0.4767	0.7870	0.7697	0.9187	1.1387
	Rank 15%	0.9087	0.9158	0.8820	0.7432	0.6100	0.7906	0.8125	0.4713	0.7846	0.7688	0.9187	1.1688
	Energy 5%	0.8788	0.9340	0.8713	0.5853	0.5920	0.7918	0.8200	0.4960	0.7528	0.7469	0.9153	0.9955
	Energy 15%	0.9326	0.9242	0.8847	0.6880	0.5980	0.8152	0.8150	0.4687	0.7809	0.7675	0.9180	1.1346
	Energy 25%	0.9198	0.9219	0.8820	0.7266	0.5940	0.8076	0.8140	0.4773	0.7870	0.7700	0.9173	1.1299
	Energy 90%	0.9053	0.9113	0.8800	0.7411	0.5900	0.7766	0.8140	0.4773	0.7834	0.7643	0.9180	1.0966
	Eff.-rank	0.9061	0.9090	0.8807	0.7350	0.5960	0.7743	0.8155	0.4773	0.7821	0.7640	0.9193	1.1167
	Base+bulk	0.8123	0.9409	0.6860	0.3806	0.5220	0.6930	0.8335	0.5347	0.6977	0.6778	0.9127	0.3994
ReClor	Base	0.8643	0.9295	0.5060	0.3460	0.5320	0.6041	**0.8464**	0.4767	0.6548	0.6400	0.5120	n/a
	Full-SFT	**0.9121**	0.8529	0.8740	0.4442	0.3580	0.7626	0.7986	0.4593	0.7234	0.6872	0.8840	n/a
	L1-reg	0.9087	0.9234	0.8773	0.3682	0.4080	0.7620	0.8193	0.4893	0.7442	**0.7000**	0.8660	0.6261
	WiSE-FT	0.9087	0.7506	**0.8780**	0.3086	0.3920	0.7556	0.8164	0.2587	0.7185	0.6430	0.8560	0.0000
	V-SoftMask	0.9053	0.8696	0.8707	0.4445	0.3420	**0.7661**	0.7955	0.4733	0.7271	0.6882	**0.8900**	0.0000
	CoFi-Tune	0.9061	**0.9363**	0.8240	0.2495	0.3860	0.6930	0.7948	**0.5047**	**0.7601**	0.6727	0.8200	0.4554
	LoRA	0.8251	0.9037	0.8220	0.2193	0.3780	0.6807	0.8007	0.4780	0.7393	0.6497	0.8080	0.3090
	FAPM	0.8677	**0.9363**	0.4993	0.3553	**0.5460**	0.6018	0.8385	0.4707	0.6499	0.6406	0.4860	0.0000
	DG-Hard (Ours)	0.9096	0.8249	0.8767	**0.4543**	0.5180	0.7591	0.8085	0.2887	0.7283	0.6853	0.8520	**0.6478**
	DG-soft	0.9087	0.7961	0.8767	0.4177	0.5260	0.7556	0.8115	0.2660	0.7234	0.6757	0.8440	0.5792
	Rank 5%	0.9104	0.8476	0.8773	0.4344	0.5360	0.7591	0.8055	0.3260	0.7319	0.6920	0.8700	0.7504
	Rank 10%	0.9113	0.8658	0.8780	0.4184	0.4880	0.7632	0.8035	0.3793	0.7258	0.6926	0.8740	0.6567
	Rank 15%	0.9130	0.8688	0.8760	0.4236	0.4540	0.7602	0.8040	0.4080	0.7283	0.6929	0.8800	0.5613
	Energy 5%	0.8814	0.9295	0.8613	0.4006	0.5480	0.7368	0.8125	0.3640	0.7381	0.6969	0.8400	0.9025
	Energy 15%	0.9087	0.8635	0.8800	0.4697	0.5440	0.7602	0.8060	0.3240	0.7319	0.6987	0.8660	0.8271
	Energy 25%	0.9113	0.8772	0.8780	0.4478	0.5300	0.7585	0.8040	0.3647	0.7271	0.6998	0.8740	0.8094
	Energy 90%	0.9138	0.8514	0.8720	0.4278	0.3660	0.7643	0.8010	0.4500	0.7258	0.6858	0.8840	0.0580
	Eff.-rank	0.9147	0.8484	0.8747	0.4365	0.3540	0.7626	0.8015	0.4573	0.7222	0.6858	0.8860	0.0000
	Base+bulk	0.9044	0.7271	0.8673	0.3541	0.5080	0.7474	0.8145	0.3520	0.7099	0.6650	0.8460	0.2308
RTE	Base	0.8643	0.9295	0.5060	0.3460	0.5320	0.6041	0.8464	0.4767	0.6548	0.6400	0.8484	n/a
	Full-SFT	0.7713	0.9242	**0.8847**	0.4150	**0.6140**	**0.7901**	0.8202	0.3000	0.7356	0.6950	**0.9495**	n/a
	L1-reg	0.7841	0.9348	0.8773	**0.4327**	0.5980	0.7538	**0.8508**	0.4780	0.7173	**0.7141**	0.9386	**0.8015**
	WiSE-FT	0.8157	0.9409	0.7773	0.3883	0.5640	0.7082	0.8455	0.4927	0.7356	0.6965	0.9386	0.7579
	V-SoftMask	0.8635	0.9348	0.7847	0.3547	0.5340	0.6766	0.8500	0.3767	**0.7491**	0.6804	0.9314	0.5982
	CoFi-Tune	0.8456	**0.9439**	0.5187	0.3488	0.4780	0.6889	0.8415	**0.5153**	0.7136	0.6549	0.9386	0.2348
	LoRA	**0.8686**	0.9318	0.4927	0.3529	0.5180	0.6088	0.8413	0.4787	0.6499	0.6381	0.9242	0.0000
	FAPM	0.8609	0.9393	0.4953	0.2288	0.3640	0.6251	0.8469	0.4760	0.6646	0.6112	0.8484	0.0000
	DG-Hard (Ours)	0.7517	0.9401	0.8793	0.4113	0.5900	0.7766	0.8425	0.4760	0.7417	0.7121	0.9314	0.7200

continued on the next page

Task	Method	ARC	GSM8K	HSwag	IFEval	Math	MMLU	MNLI	Trivia	TruthQA	Avg.	Results	HM(h, p)
	DG-soft	0.7500	0.9439	0.8700	0.3984	0.5900	0.7778	0.8420	0.4853	0.7417	0.7110	0.9206	0.7332
	Rank 5%	0.7602	0.9409	0.8833	0.4258	0.6040	0.7936	0.8390	0.4407	0.7295	0.7130	0.9350	0.6483
	Rank 10%	0.7602	0.9371	0.8827	0.4200	0.5920	0.7918	0.8385	0.4053	0.7344	0.7069	0.9422	0.5148
	Rank 15%	0.7654	0.9348	0.8840	0.4325	0.5980	0.8006	0.8345	0.3693	0.7344	0.7059	0.9422	0.3814
	Energy 5%	0.8626	0.9477	0.5347	0.3619	0.5600	0.6994	0.8080	0.5100	0.7246	0.6677	0.8773	0.4710
	Energy 15%	0.7517	0.9454	0.8613	0.4000	0.5960	0.7626	0.8425	0.4887	0.7442	0.7103	0.9278	0.7408
	Energy 25%	0.7398	0.9454	0.8887	0.4254	0.6020	0.8047	0.8385	0.4473	0.7307	0.7136	0.9350	0.6034
	Energy 90%	0.7688	0.9265	0.8860	0.4120	0.6280	0.7942	0.8265	0.3040	0.7319	0.6975	0.9495	0.0106
	Eff.-rank	0.7705	0.9249	0.8860	0.4120	0.6160	0.7901	0.8240	0.3027	0.7368	0.6959	0.9495	0.0134
	Base+bulk	0.8780	0.9371	0.5287	0.3660	0.5440	0.6778	0.8490	0.4213	0.7222	0.6582	0.9350	0.3807

Table 23: Per-benchmark held-out scores for Llama-3.2-3B-Instruct across all (task, method) pairs, with non-spectral methods above and spectral rules below the rule in each task block. **Bold** = best per column within the task block; underline = second-best (gap ≥ 0.001 from best). Avg. = mean of the nine held-out benchmarks. Results = on-task `task_{name}` score. HM(h, p) $=$ HM(% healed, % preserved) on the partitioned held-out set of the cell, each a ratio of sums as in §3.4, floored at 0 and otherwise unclipped. Cells with no damaged or no improved benchmarks default the corresponding side to 100. Base and Full-SFT have no defined HM(h, p). Bold/underline on it follow the same convention as other columns, restricted to the seven repair methods (excluding the two reference rows). "Repair methods" here covers both training-time interventions (L1-reg, V-SoftMask (Ke et al., 2023), CoFi-Tune (Zhang et al., 2024a), LoRA (Hu et al., 2022)) and post-hoc methods (WiSE-FT (Wortsman et al., 2022), FAPM (Huang et al., 2025), DG-Hard). See §4.1 for this grouping of the methods. Spectral rules (§4.4) are listed for reference and are excluded from bold/underline. Their fractions are of the rank or of the spectral energy, and rank 10% is the TSV-C setting.

Task	Method	ARC	GSM8K	HSwag	IFEval	Math	MMLU	MNLI	Trivia	TruthQA	Avg.	Results	HM(h, p)
MedQA	Base	0.7543	0.7218	0.6193	**0.7784**	0.4200	0.3474	0.5025	0.3360	0.5606	0.5600	0.4949	n/a
	Full-SFT	0.7440	0.7058	0.6980	0.7599	0.4180	0.5655	0.4665	**0.5253**	0.6157	0.6110	0.5774	n/a
	L1-reg	0.7457	0.7225	0.6787	0.7722	0.3980	0.4094	0.4745	0.4367	0.5973	0.5817	0.5695	0.3034
	WiSE-FT	0.7517	0.7202	0.6727	0.7667	0.4120	0.3860	0.4710	0.4267	0.5961	0.5781	**0.5829**	0.1908
	V-SoftMask	0.7474	0.7096	**0.7007**	0.7557	0.4360	**0.5719**	0.4610	0.5247	0.6108	**0.6131**	0.5742	0.0000
	CoFi-Tune	0.7534	0.7134	0.6460	0.7740	**0.4620**	0.3585	0.4415	0.3900	0.6059	0.5716	0.5687	0.0000
	LoRA	0.7483	**0.7339**	0.6440	0.7612	0.4300	0.3953	0.4740	0.4027	**0.6193**	0.5788	0.5734	0.2655
	FAPM	**0.7594**	0.7218	0.6180	0.7766	0.4220	0.3538	0.5005	0.3427	0.5545	0.5610	0.5035	0.0206
	DG-Hard (Ours)	0.7449	0.7301	0.6693	0.7666	0.4280	0.3848	**0.5045**	0.4113	0.5924	0.5813	0.5774	**0.5364**
	DG-soft	0.7568	0.7180	0.6627	0.7760	0.4160	0.3585	0.4855	0.3973	0.5985	0.5744	0.5782	0.3693
	Rank 5%	0.7466	0.7142	0.6840	0.7673	0.4220	0.3953	0.4730	0.4493	0.5973	0.5832	0.5782	0.2632
	Rank 10%	0.7389	0.6983	0.6847	0.7590	0.4240	0.4181	0.4705	0.4560	0.6071	0.5841	0.5789	0.1854
	Rank 15%	0.7372	0.7111	0.6900	0.7709	0.4240	0.4368	0.4695	0.4700	0.6071	0.5907	0.5742	0.1472
	Energy 5%	0.7594	0.7445	0.6613	0.7646	0.4220	0.3602	0.4875	0.4120	0.6010	0.5792	0.5742	0.4103
	Energy 15%	0.7526	0.7301	0.6753	0.7550	0.4260	0.3819	0.4740	0.4353	0.6034	0.5815	0.5719	0.2807
	Energy 25%	0.7449	0.7104	0.6887	0.7582	0.4320	0.4158	0.4695	0.4547	0.6034	0.5864	0.5766	0.1448
	Energy 90%	0.7423	0.6960	0.7020	0.7685	0.4420	0.5158	0.4535	0.5073	0.6132	0.6045	0.5805	0.0000
	Eff.-rank	0.7440	0.6922	0.7013	0.7676	0.4260	0.5363	0.4520	0.5180	0.6144	0.6058	0.5797	0.0000
	Base+bulk	0.7568	0.7149	0.6773	0.7636	0.4200	0.3690	0.4320	0.4433	0.6083	0.5761	0.5797	0.0000
WikiQA	Base	0.7543	0.7218	**0.6193**	0.7784	0.4200	0.3474	0.5025	0.3360	0.5606	0.5600	0.9526	n/a
	Full-SFT	0.7474	0.6520	0.5947	0.7354	0.4320	0.3404	0.5580	**0.4933**	**0.6108**	0.5738	**0.9664**	n/a
	L1-reg	0.7389	0.7210	0.5347	0.7703	0.4200	0.3404	0.4920	0.4320	0.5655	0.5572	0.9638	0.5006
	WiSE-FT	0.7423	0.6808	0.5947	0.7535	0.4340	0.3491	0.5425	0.4547	0.5936	0.5717	0.9629	0.5292
	V-SoftMask	0.7381	0.6368	0.5920	0.7423	**0.4380**	0.3392	**0.5600**	0.4900	0.6083	0.5716	0.9663	0.0000
	CoFi-Tune	0.7193	0.6854	0.5687	**0.7825**	0.4200	0.3058	0.4790	0.3813	0.6095	0.5502	0.9658	0.3908
	LoRA	0.7329	0.7180	0.5687	0.7644	0.4360	0.3450	0.4705	0.4380	0.5802	0.5615	0.9646	0.4850
	FAPM	**0.7688**	0.7225	0.6127	0.7743	0.4320	**0.3532**	0.5050	0.3407	0.5557	0.5628	0.9531	0.0171
	DG-Hard (Ours)	0.7415	**0.7293**	0.5833	0.7720	0.4220	0.3433	0.5535	0.4833	0.5961	**0.5805**	0.9580	**0.9456**
	DG-soft	0.7491	0.7384	0.5847	0.7591	0.4100	0.3456	0.5295	0.4580	0.5777	0.5725	0.9573	0.7671
	Rank 5%	0.7483	0.7437	0.5753	0.7748	0.4120	0.3485	0.5420	0.4673	0.6120	0.5804	0.9613	0.9788
	Rank 10%	0.7577	0.7377	0.5800	0.7599	0.4080	0.3480	0.5415	0.4753	0.6181	0.5807	0.9620	0.9350
	Rank 15%	0.7611	0.7263	0.5833	0.7553	0.4060	0.3450	0.5425	0.4780	0.6316	0.5810	0.9627	0.8942
	Energy 5%	0.7526	0.7437	0.5740	0.7693	0.4260	0.3526	0.5375	0.4420	0.5887	0.5763	0.9567	0.8154
	Energy 15%	0.7449	0.7384	0.5753	0.7753	0.4380	0.3515	0.5490	0.4667	0.6144	0.5837	0.9600	0.9845
	Energy 25%	0.7551	0.7407	0.5733	0.7489	0.4260	0.3480	0.5420	0.4673	0.6206	0.5802	0.9613	0.8918
	Energy 90%	0.7440	0.7218	0.5813	0.7378	0.4020	0.3427	0.5460	0.4773	0.6230	0.5751	0.9647	0.7615
	Eff.-rank	0.7423	0.7225	0.5860	0.7408	0.4200	0.3339	0.5540	0.4740	0.6267	0.5778	0.9660	0.7950
	Base+bulk	0.7526	0.7271	0.6013	0.7702	0.4320	0.3456	0.5120	0.4433	0.5863	0.5745	0.9593	0.6964
Winogrande	Base	0.7543	0.7218	0.6193	0.7784	0.4200	0.3474	0.5025	0.3360	0.5606	0.5600	0.5209	n/a
	Full-SFT	0.7765	0.6960	**0.6673**	0.7647	0.4140	0.4275	0.5370	0.4533	0.4835	0.5800	**0.8437**	n/a
	L1-reg	0.7167	0.7134	0.5947	0.7524	0.4120	0.3222	0.5150	0.4487	0.5422	0.5575	0.7956	0.3978
	WiSE-FT	0.7671	0.6967	0.6547	0.7715	**0.4320**	0.3731	0.5375	0.4627	0.5483	0.5826	0.7719	0.8178
	V-SoftMask	**0.7782**	0.7058	**0.6673**	0.7572	0.4100	**0.4351**	0.5365	0.4507	0.4786	0.5799	0.8421	0.0000
	CoFi-Tune	0.7218	0.6960	0.6193	0.7737	**0.4320**	0.3427	**0.5765**	0.3920	**0.5692**	0.5692	0.7301	0.6382
	LoRA	0.7560	0.6922	0.6173	**0.7874**	**0.4320**	0.3719	0.4900	0.3707	0.4933	0.5567	0.8114	0.1415
	FAPM	0.7585	0.7233	0.6160	0.7800	0.4240	0.3515	0.5035	0.3393	0.5557	0.5613	0.5178	0.0357
	DG-Hard (Ours)	0.7517	**0.7339**	0.6580	0.7713	0.4060	0.3854	0.5475	**0.4680**	0.5447	**0.5852**	0.8011	**0.8462**

continued on the next page

Task	Method	ARC	GSM8K	HSwag	IFEval	Math	MMLU	MNLI	Trivia	TruthQA	Avg.	Results	HM(h, p)
	DG-soft	0.7637	0.7377	0.6540	0.7675	0.4180	0.3579	0.5390	0.4787	0.5581	0.5861	0.7790	0.8770
	Rank 5%	0.7739	0.7362	0.6653	0.7768	0.4080	0.3836	0.5405	0.4493	0.5288	0.5847	0.8279	0.6894
	Rank 10%	0.7747	0.7384	0.6713	0.7713	0.4060	0.3918	0.5400	0.4393	0.5141	0.5830	0.8319	0.5406
	Rank 15%	0.7739	0.7271	0.6720	0.7684	0.4120	0.4012	0.5355	0.4347	0.5165	0.5824	0.8335	0.5700
	Energy 5%	0.7662	0.7483	0.6453	0.7801	0.4240	0.3409	0.5460	0.4847	0.5594	0.5883	0.7735	0.8553
	Energy 15%	0.7722	0.7309	0.6600	0.7760	0.3960	0.3836	0.5455	0.4640	0.5226	0.5834	0.8287	0.6456
	Energy 25%	0.7730	0.7339	0.6700	0.7837	0.4220	0.3901	0.5410	0.4473	0.5239	0.5872	0.8335	0.6535
	Energy 90%	0.7773	0.7301	0.6687	0.7779	0.4160	0.4152	0.5350	0.4400	0.5006	0.5845	0.8414	0.3569
	Eff.-rank	0.7756	0.7324	0.6673	0.7644	0.4160	0.4246	0.5340	0.4440	0.4884	0.5830	0.8429	0.1190
	Base+bulk	0.7363	0.7210	0.6493	0.7761	0.4340	0.3363	0.5315	0.4240	0.5716	0.5756	0.7340	0.6814
StrategyQA	Base	0.7543	0.7218	0.6193	**0.7784**	0.4200	0.3474	0.5025	0.3360	0.5606	0.5600	0.6521	n/a
	Full-SFT	0.7457	0.5519	**0.6220**	0.6854	0.3720	**0.5228**	0.5235	**0.5427**	0.5936	0.5733	0.7220	n/a
	L1-reg	0.7398	0.6740	0.6040	0.7228	0.4000	0.4766	0.5130	0.5213	0.6010	0.5836	**0.7234**	0.7073
	WiSE-FT	0.7363	0.6998	0.6100	0.7334	0.4380	0.4211	0.5190	0.4940	**0.6083**	**0.5844**	0.7074	**0.7483**
	V-SoftMask	0.7526	0.5557	0.6173	0.6808	0.3960	0.5187	**0.5320**	0.5393	0.5924	0.5761	0.7176	0.1385
	CoFi-Tune	0.7227	0.6884	0.5753	0.7708	0.4200	0.3491	0.5165	0.4427	0.6059	0.5657	0.6929	0.5191
	LoRA	0.7389	**0.7240**	0.6020	0.7730	**0.4400**	0.3164	0.5070	0.3687	0.5594	0.5588	0.6623	0.0022
	FAPM	**0.7551**	0.7187	0.6153	0.7729	0.4160	0.3485	0.5050	0.3367	0.5508	0.5577	0.6463	0.0000
	DG-Hard (Ours)	0.7270	0.7058	0.6080	0.7324	**0.4400**	0.3772	0.5250	0.4767	0.5973	0.5766	0.6914	0.6330
	DG-soft	0.7355	0.7240	0.5980	0.7607	0.4300	0.3480	0.5085	0.4780	0.6083	0.5768	0.6827	0.6252
	Rank 5%	0.7466	0.7255	0.6027	0.7407	0.4200	0.3819	0.5075	0.5060	0.6181	0.5832	0.7060	0.7388
	Rank 10%	0.7415	0.7263	0.5993	0.6982	0.4140	0.4099	0.5115	0.5147	0.6108	0.5807	0.7016	0.7193
	Rank 15%	0.7483	0.7271	0.6027	0.7067	0.3980	0.4263	0.5080	0.5167	0.6120	0.5829	0.7118	0.7320
	Energy 5%	0.7261	0.7377	0.5927	0.7652	0.4340	0.3257	0.5060	0.4260	0.5802	0.5659	0.6696	0.3528
	Energy 15%	0.7321	0.7286	0.6027	0.7557	0.4200	0.3737	0.5115	0.4967	0.6132	0.5816	0.7045	0.7178
	Energy 25%	0.7440	0.7218	0.6053	0.7324	0.4260	0.3977	0.5130	0.5093	0.6157	0.5850	0.7016	0.7583
	Energy 90%	0.7491	0.6793	0.6073	0.6815	0.4020	0.4918	0.5120	0.5393	0.5998	0.5847	0.7234	0.6454
	Eff.-rank	0.7466	0.6763	0.6093	0.6710	0.3900	0.5012	0.5115	0.5420	0.6047	0.5836	0.7191	0.5784
	Base+bulk	0.7381	0.7210	0.5973	0.7163	0.4100	0.4146	0.4960	0.5107	0.6218	0.5806	0.7074	0.7475
BoolQ	Base	0.7543	0.7218	0.6193	**0.7784**	0.4200	0.3474	0.5025	0.3360	0.5606	0.5600	0.7859	n/a
	Full-SFT	**0.7628**	0.6876	**0.6353**	0.7187	0.4040	**0.4801**	0.4900	**0.4967**	0.5679	**0.5826**	0.8902	n/a
	L1-reg	0.7543	0.7233	0.5520	0.7756	0.4160	0.3661	0.4845	0.4807	0.5887	0.5712	0.8872	**0.7119**
	WiSE-FT	0.7551	0.7240	0.6120	0.7690	0.4240	0.3813	0.4995	0.4647	0.5875	0.5797	0.8731	0.6928
	V-SoftMask	0.7517	0.6657	0.6067	0.7046	**0.4440**	0.4520	0.4715	0.4847	0.5875	0.5743	**0.8951**	0.0000
	CoFi-Tune	0.7346	0.7119	0.5687	0.7681	0.4340	0.3105	0.4555	0.4313	**0.5985**	0.5570	0.8832	0.3180
	LoRA	0.7423	**0.7377**	0.5607	0.7606	0.4400	0.3298	0.4685	0.4567	0.5789	0.5639	0.8743	0.5174
	FAPM	0.7551	0.7149	0.6167	0.7693	0.4140	0.3538	0.5000	0.3387	0.5532	0.5573	0.7856	0.0598
	DG-Hard (Ours)	0.7526	0.7286	0.6147	0.7686	0.4280	0.3673	**0.5065**	0.4720	0.5802	0.5798	0.8633	0.6859
	DG-soft	0.7671	0.7400	0.5987	0.7585	0.4260	0.3485	0.4920	0.4527	0.5826	0.5740	0.8567	0.5700
	Rank 5%	0.7517	0.7384	0.6067	0.7545	0.4360	0.3573	0.4930	0.4647	0.5814	0.5760	0.8787	0.6250
	Rank 10%	0.7543	0.7187	0.6093	0.7475	0.4160	0.3731	0.4930	0.4687	0.5851	0.5740	0.8827	0.5848
	Rank 15%	0.7526	0.7278	0.6187	0.7272	0.4240	0.3825	0.4895	0.4713	0.5789	0.5747	0.8840	0.5482
	Energy 5%	0.7602	0.7460	0.5993	0.7764	0.4340	0.3649	0.4960	0.4573	0.5851	0.5799	0.8627	0.6847
	Energy 15%	0.7423	0.7339	0.6073	0.7666	0.4160	0.3696	0.4935	0.4720	0.5838	0.5761	0.8773	0.7016
	Energy 25%	0.7543	0.7309	0.6100	0.7636	0.4320	0.3772	0.4930	0.4620	0.5851	0.5787	0.8820	0.6785
	Energy 90%	0.7517	0.7187	0.6287	0.7167	0.4260	0.4164	0.4895	0.4867	0.5777	0.5791	0.8880	0.4382
	Eff.-rank	0.7551	0.7202	0.6287	0.6999	0.4140	0.4345	0.4890	0.4947	0.5741	0.5789	0.8900	0.2502
	Base+bulk	0.7543	0.7157	0.6107	0.7599	0.4060	0.3632	0.4835	0.4600	0.6047	0.5731	0.8867	0.5792
ReClor	Base	0.7543	0.7218	0.6193	0.7784	0.4200	0.3474	0.5025	0.3360	0.5606	0.5600	0.4860	n/a
	Full-SFT	0.7628	0.7468	**0.6573**	0.7558	0.4300	**0.4386**	0.5290	**0.4860**	**0.6255**	**0.6035**	**0.6800**	n/a
	L1-reg	0.7526	0.7354	0.6420	0.7548	0.4460	0.3731	**0.5330**	0.4427	0.6181	0.5886	0.6760	0.7638
	WiSE-FT	**0.7688**	0.7263	0.6373	**0.7856**	**0.4500**	0.3532	0.5155	0.4467	0.6083	0.5880	0.6180	0.6925
	V-SoftMask	0.7628	**0.7521**	**0.6573**	0.7492	0.4320	0.4310	0.5260	0.4780	0.6230	0.6013	**0.6800**	**0.9731**
	CoFi-Tune	0.7355	0.7233	0.6053	0.7689	0.4360	0.3310	0.5260	0.3907	0.6157	0.5703	0.5960	0.3749
	LoRA	0.7415	0.7233	0.6140	0.7764	0.4420	0.3257	0.5155	0.3640	0.5887	0.5657	0.5620	0.1563
	FAPM	0.7560	0.7187	0.6187	0.7740	0.4240	0.3596	0.5030	0.3453	0.5508	0.5611	0.4900	0.0628
	DG-Hard (Ours)	0.7560	0.7324	0.6340	0.7733	0.4420	0.3386	0.5255	0.4567	0.6059	0.5849	0.6040	0.6661
	DG-soft	0.7602	0.7468	0.6373	0.7732	0.4080	0.3310	0.5150	0.4500	0.6108	0.5814	0.5960	0.6503
	Rank 5%	0.7713	0.7475	0.6360	0.7714	0.4220	0.3450	0.5150	0.4513	0.6255	0.5872	0.6200	0.7223
	Rank 10%	0.7739	0.7528	0.6420	0.7607	0.4440	0.3702	0.5145	0.4507	0.6230	0.5924	0.6420	0.7855
	Rank 15%	0.7696	0.7483	0.6453	0.7559	0.4280	0.3860	0.5210	0.4460	0.6279	0.5920	0.6560	0.8256
	Energy 5%	0.7560	0.7551	0.6293	0.7672	0.4120	0.3269	0.5200	0.4460	0.6181	0.5812	0.5920	0.6268
	Energy 15%	0.7739	0.7392	0.6373	0.7706	0.4200	0.3462	0.5180	0.4480	0.6218	0.5861	0.6140	0.7116
	Energy 25%	0.7577	0.7483	0.6400	0.7741	0.4200	0.3690	0.5245	0.4453	0.6255	0.5894	0.6320	0.7724
	Energy 90%	0.7628	0.7574	0.6560	0.7619	0.4380	0.4193	0.5225	0.4607	0.6255	0.6005	0.6740	0.9284
	Eff.-rank	0.7628	0.7589	0.6607	0.7467	0.4380	0.4281	0.5260	0.4660	0.6279	0.6017	0.6760	0.9627
	Base+bulk	0.7713	0.7377	0.6500	0.7706	0.4240	0.3608	0.4950	0.4433	0.6083	0.5846	0.6360	0.7333

continued on the next page

Task	Method	ARC	GSM8K	HSwag	IFEval	Math	MMLU	MNLI	Trivia	TruthQA	Avg.	Results	HM(h, p)
RTE	Base	0.7543	0.7218	**0.6193**	0.7784	0.4200	0.3474	0.5025	0.3360	0.5606	0.5600	0.5957	n/a
	Full-SFT	0.7346	0.6308	0.6013	0.7031	0.3660	**0.5228**	0.5870	**0.5053**	0.5802	0.5812	**0.9061**	n/a
	L1-reg	0.7082	0.6960	0.5847	0.7508	0.3980	0.4398	0.5895	0.4693	**0.6034**	0.5822	0.8845	**0.6914**
	WiSE-FT	0.7142	0.7051	0.5853	0.7686	0.4180	0.3889	0.5630	0.4620	0.6010	0.5784	0.8773	0.6598
	V-SoftMask	0.7321	0.6240	0.6087	0.7003	0.3800	0.5187	**0.6050**	**0.5053**	0.5924	**0.5852**	**0.9061**	0.0392
	CoFi-Tune	0.7014	0.6808	0.5600	0.7534	0.4280	0.3205	0.5225	0.4160	0.6010	0.5537	0.8736	0.2766
	LoRA	0.7321	**0.7309**	0.6053	**0.7860**	0.4220	0.3392	0.5320	0.3727	0.5753	0.5662	0.8448	0.2402
	FAPM	**0.7568**	0.7165	0.6127	0.7750	0.4080	0.3520	0.5045	0.3407	0.5618	0.5587	0.5993	0.0514
	DG-Hard (Ours)	0.7201	0.7081	0.5840	0.7726	**0.4300**	0.3667	0.5620	0.4467	**0.6034**	0.5771	0.8592	0.6041
	DG-soft	0.7244	0.7134	0.5747	0.7705	0.4040	0.3345	0.5380	0.4427	0.5985	0.5667	0.8484	0.4453
	Rank 5%	0.7159	0.7036	0.5740	0.7609	0.4040	0.3684	0.5550	0.4773	0.6144	0.5748	0.8736	0.6053
	Rank 10%	0.7210	0.6967	0.5787	0.7418	0.3920	0.4047	0.5625	0.4827	0.6193	0.5777	0.8773	0.6039
	Rank 15%	0.7133	0.6937	0.5700	0.7425	0.4180	0.4263	0.5635	0.4880	0.6132	0.5810	0.8773	0.6902
	Energy 5%	0.7346	0.7309	0.5820	0.7648	0.4360	0.3275	0.5300	0.3993	0.5643	0.5633	0.8448	0.2857
	Energy 15%	0.7116	0.7111	0.5780	0.7627	0.4080	0.3602	0.5550	0.4587	0.6071	0.5725	0.8628	0.5725
	Energy 25%	0.7133	0.7036	0.5833	0.7500	0.3980	0.3930	0.5580	0.4760	0.6181	0.5770	0.8773	0.6187
	Energy 90%	0.7287	0.6657	0.5907	0.7180	0.3820	0.4918	0.5760	0.5027	0.5924	0.5831	0.8989	0.4477
	Eff.-rank	0.7278	0.6664	0.5927	0.7109	0.3760	0.5070	0.5745	0.5073	0.5949	0.5842	0.8953	0.3854
	Base+bulk	0.7108	0.6998	0.5767	0.7649	0.3940	0.3655	0.5590	0.4833	0.6047	0.5732	0.8809	0.6021

K.1 METHOD-SPECIFIC CONFIGURATIONS

L1-reg adds an L1 penalty $\|W - W_{\text{base}}\|_1$ to the loss with $\lambda = 1.0 \times 10^{-6}$. **LoRA** uses rank $r = 16$, $\alpha = 32$, dropout 0.05, and auto-selected target modules. **WiSE-FT** is applied post-hoc to the Full-SFT checkpoint with mixing weight $\alpha = 0.5$, which gives the linear weight average $W^* = \frac{1}{2}W_{\text{base}} + \frac{1}{2}W_{\text{ft}}$. This is the canonical setting from Wortsman et al. (2022). **FAPM** reverts the entries of Δ ranked lowest under the forgetting-aware FAPM score, which combines absolute change with a relative-change penalty against W_{base}. We use the published 90% reversion rate. **V-SoftMask** (our reproduction of DAS, Ke et al., 2023) uses a calibration batch size of 4 and applies mask $= (1 -$ normalized gradient importance$)$, fixed for the SFT run. **CoFi-Tune** uses the two-stage configuration of Zhang et al. (2024a) verbatim. We run the coarse layer-range filter at the second-quartile setting $(N \times 25\%, N \times 50\%]$. **DG-Hard** applies the $\lambda^*(\beta)\,\hat{\sigma}\sqrt{\max(m,n)}$ hard threshold of §3. It has no tuning knobs and uses no calibration data. **TIES** (Yadav et al., 2023) keeps the top-k entries of $|\Delta|$ by magnitude, sign-elects, and averages. We use `keep_ratio` $= 0.2$ ($k = 20\%$) with mixing weight $\lambda = 1$. **DARE** (Yu et al., 2024) drops entries of Δ with probability p and rescales the survivors by $1/(1-p)$. We use `drop_prob` $= 0.5$ with the rescaled variant. TIES and DARE are evaluated only in the App. L.3 ablation, and are absent from the headline comparison of Tab. 22 and Tab. 23.

Spectral controls (§4.4). Every control takes the same per-matrix SVD as DG-Hard and differs only in the rule that selects the kept singular values. **Fixed-rank** keeps the top $\text{round}(f \cdot \min(m,n))$ components for $f \in \{0.05, 0.10, 0.15\}$. The value $f = 0.10$ is the rank fraction at which we evaluate TSV-C (Gargiulo et al., 2025) (§2). **Fixed energy-percentage** keeps the fewest leading components whose cumulative squared singular values reach a fraction $e \in \{0.05, 0.15, 0.25, 0.90\}$ of $\sum_i s_i^2$. **Effective-rank** keeps the top $k = \text{round}(\exp(H))$ components with $H = -\sum_i p_i \log p_i$ and $p_i = s_i / \sum_j s_j$ (Roy & Vetterli, 2007). **DG-soft** applies the Frobenius-optimal shrinker of Gavish & Donoho (2017) with the same noise estimate $\hat{\sigma}$ as DG-Hard. It zeroes singular values below the Marchenko-Pastur edge and shrinks every surviving one. **Base + noise-bulk** reconstructs W_{base} plus the components that DG-Hard discards, that is, $\Delta - \Delta^*$. None of the controls uses data at repair time; selecting a fraction does (§4.4). The fixed-rank and energy-percentage fractions are hyperparameters, and the reported values are the ones evaluated.

K.2 INFERENCE ENGINE AND DECODING

All inference uses vLLM (Kwon et al., 2023) with `-enforce-eager`, `max_model_len=8192`, `dtype=bfloat16`, greedy decoding (temperature $= 0$), *no* `max_tokens` cap, and the model's native chat template applied via the tokenizer's `apply_chat_template` method. Greedy decoding under vLLM's continuous batching is not bit-deterministic, so the same prompt can produce slightly different completions on re-runs. We measure a **drift floor of about ± 1 pp per benchmark**. We treat any delta smaller than this floor as noise and use a **3 pp significance threshold** throughout the paper. The threshold is conservative ($\approx 3\times$ the drift floor) and is chosen to capture genuine but moderate

fine-tuning effects while excluding sampling drift. Output scoring strips <think>... </think> blocks and reasoning preludes before metric computation.

L Other experiments and ablations

This appendix collects the experiments that support the design choices of DG-Hard. App. L.1 quantifies which weight classes carry the noise mass that DG-Hard removes (layer-class noise concentration). App. L.2 tests the layer-mask choice. App. L.3 compares DG-Hard against element-wise merging baselines at canonical and matched-rollback operating points. App. M asks what happens when the task signal is diffuse, and App. N positions the method against data-based repair and unlearning.

L.1 Layer-wise noise concentration

For each (model, task) fine-tune, we cache the SVD of every 2D weight delta and classify each matrix by its layer class (e.g., `mlp.up_proj`, `attn.q_proj`). For each class we report the mean noise-energy fraction (the share of singular-value-squared mass that lies below the MP edge λ_{MP} of (3)) and the total norm-share (the fraction of the network's Frobenius^2 delta mass attributable to the class). Their product is the operationally meaningful noise mass per class under the signal-plus-noise model of (2). Averaged across six (model, task) pairs, the MLP gate and up projections carry most of the noise mass that DG-Hard removes (Tab. 24). Attention and embedding classes contribute substantially less.

Table 24: Per-class noise ranking, averaged over six (model, task) pairs. Mean noise is the fraction of squared singular-value mass below λ_{MP}. Mean norm-share is the class share of the network's total Frobenius^2 delta. Their product is the noise mass per class. Top-two rows in bold.

Class	Mean noise	Mean norm-share	Noise × norm-share
`mlp.up_proj`	**0.93**	**0.231**	**0.213**
`mlp.gate_proj`	**0.93**	**0.214**	**0.196**
`mlp.down_proj`	0.55	0.238	0.131
`attn.v_proj`	0.41	0.045	0.018
`attn.o_proj`	0.10	0.099	0.010
`attn.k_proj`	0.37	0.022	0.008
`embed_tokens`	≈1.00	0.026	0.026

Metric note. The mean-noise column above (energy fraction of squared singular values below λ_{MP}) and the MP-fit score of (8) (a Kolmogorov-Smirnov distance between the empirical and theoretical CDFs) measure different things. The former is the share of squared mass under the bulk edge. The latter is how closely the spectrum's shape matches the MP density. The closeness of the `mlp.up_proj` row's 0.93 to the 93.1% MP-fit reported in App. A is therefore a coincidence between two different quantities. The `embed_tokens` value ≈ 1.00 is a degenerate case in which essentially all of the FT delta's spectral mass sits under the bulk edge. The embedding update therefore carries no above-edge spike on these (model, task) pairs. This is consistent with the claim of App. A that the learned signal is concentrated in a small number of spike directions, which lie outside the embedding for these fine-tunes.

The aggregate ranking conceals a sharp cross-family asymmetry in attention noise (Tab. 25). On Llama-3.2-3B, `attn.q_proj` has mean below-edge mass 0.025, while on Qwen3.5-4B the same class shows 0.91, a $36\times$ ratio. The `attn.o_proj` class shows a $7\times$ ratio in the same direction. MLP classes are family-invariant (ratios within 1.16).

The family asymmetry is consistent with the hypothesis that reasoning post-training (long chain-of-thought RL or distillation, present in Qwen3.5-4B but not in Llama-3.2-3B-Instruct) disperses attention weights and leaves more below-edge mass in the FT delta. This explanation is, however, preliminary and untested in the present run. Whatever the underlying cause, the gate and up MLP classes carry the dominant share of removable noise mass in both families. App. L.2 tests whether restricting the repair to these classes helps.

Table 25: Cross-family asymmetry in mean noise-energy fraction. Attention classes diverge sharply between Llama-3.2-3B-Instruct and Qwen3.5-4B. MLP classes do not.

Class	Llama mean noise	Qwen mean noise	Ratio
`mlp.gate_proj`	0.93	0.94	1.01
`mlp.up_proj`	0.92	0.93	1.01
`mlp.down_proj`	0.51	0.59	1.16
`attn.q_proj`	0.025	0.91	36×
`attn.o_proj`	0.027	0.19	7×
`attn.k_proj`	0.36	0.39	1.08
`attn.v_proj`	0.44	0.35	0.80

L.2 LAYER-MASK CAUSAL TEST

App. L.1 ranks 2D weight classes by noise mass and identifies `gate_proj` and `up_proj` as the top-two carriers, averaged over the six (model, task) pairs. This appendix verifies that ranking causally. We vary the set of layers pruned by DG-Hard and measure cross-domain accuracy in a separate evaluation run on five held-out benchmarks (ARC-C, GSM8K, HellaSwag, IFEval, MMLU). The Base and FT scores and damaged sets of this run differ slightly from those of the headline run. The protocol applies DG-Hard canonical (the threshold of Algorithm 1, no scaling) to two pairs, Qwen3.5-4B with Winogrande (severe FT damage) and Llama-3.2-3B with MedQA (mild FT damage). Each pair is repaired under four masks: `ALL` (every 2D matrix), `mlp_only` (`\.mlp\.`), `attn_only` (`\.self_attn\.`), and `gate_up` (`\.mlp\.(gate|up)_proj`). Tensors outside the active mask keep the full FT delta (no shrinkage). We report Frobenius retention r alongside benchmark accuracies and the protocol-aligned Combined score. The four masks above target attention/MLP partitions of the decoder. The `embed_tokens` and `lm_head` matrices are in scope for the headline DG-Hard configuration (App. D.5), and in all four conditions of this ablation they are kept at their `ALL`-mask repaired state. This experiment focuses on the attention-vs-MLP partition. The embedding contribution is characterized separately in App. L.1.

Column conventions. `task_win` and `task_med` are on-task accuracy on Winogrande and MedQA, respectively. The Combined column applies the partition-conditional harmonic mean of (7d). When $U = \emptyset$ (Tab. 26), the non-damage term of (7) is undefined ($0/0$), and we adopt the convention non-damage $= 100$, which matches the implementation. With this convention, the `attn_only` row of Tab. 26 collapses to Combined $= 0.0$ because percent-healed is negative on the damaged benches (GSM8K, ARC-C). The harmonic mean then goes to zero independently of the U-term.

Table 26: Mask ablation on Qwen3.5-4B + Winogrande (severe FT damage). $D = \{\text{arc_challenge}, \text{gsm8k}\}$, $I = \{\text{hellaswag}, \text{ifeval}, \text{mmlu}\}$, $U = \emptyset$. Combined uses the protocol harmonic mean over the partition. Winning mask on Combined in bold.

Mask	GSM8K	ARC-C	HSwag	IFEval	MMLU	task_win	r	Combined
ALL	87.7	73.9	85.0	36.1	75.7	89.4	0.314	86.2
mlp_only	89.2	72.6	86.7	35.7	78.2	89.0	0.633	**87.8**
attn_only	58.5	55.3	87.2	41.2	75.0	89.7	0.966	0.0
gate_up	87.6	67.7	87.0	39.4	77.9	89.6	0.780	84.1

Pattern A (Qwen, severe FT). `ALL`, `mlp_only`, and `gate_up` sit within a few points of one another on GSM8K (87.6–89.2) and within seven on ARC-C (67.7–73.9). `attn_only` collapses on the damaged benches, with GSM8K dropping to 58.5 (vs. 87.7 under `ALL`) and ARC-C to 55.3 (vs. 73.9). The protocol-aligned Combined column is more discriminating still. On Combined, `attn_only` scores 0.0 because it has negative percent healed on the damaged benches, which the harmonic mean penalizes. Repairing only attention while leaving the MLP delta unshrunken keeps

Table 27: Mask ablation on Llama-3.2-3B + MedQA (mild FT damage). $D = \{\text{gsm8k}\}$, $I = \{\text{hellaswag}, \text{mmlu}\}$, $U = \{\text{arc_challenge}, \text{ifeval}\}$. Winning mask on Combined in bold.

Mask	GSM8K	ARC-C	HSwag	IFEval	MMLU	task_med	r	Combined
ALL	73.8	74.6	67.1	75.9	38.7	57.7	0.186	71.8
mlp_only	71.0	74.3	68.1	75.4	43.0	57.3	0.540	76.1
attn_only	70.4	73.9	69.6	75.9	49.2	57.7	0.874	**81.6**
gate_up	69.7	74.4	69.5	76.9	49.1	57.5	0.730	74.2

too much noise in the model, regardless of how aggressively attention is repaired. On this pair, repairing the MLP class is necessary and repairing only attention is insufficient.

Pattern B (Llama, mild FT). The four masks stay within a few points of one another on most per-benchmark accuracies (MMLU is the exception, spanning 38.7 to 49.2), and Combined separates them. With a mild FT delta, the absolute amount of noise to remove is small, so any reasonable repair yields similar cross-domain accuracy on most benchmarks. Combined, however, spreads from 71.8 (`ALL`) to 81.6 (`attn_only`) because of the MMLU and on-task differences. We treat this as a *negative* result for mask selection on a mildly damaged pair. It is not evidence that all masks are equal. The cross-family contrast is informative. On Llama, the attention q and o projections are low-noise (App. L.1), so `attn_only` removes little of the delta ($r = 0.874$). On Qwen, `attn.q_proj` has a below-edge noise fraction of 0.91 (vs. 0.025 on Llama), so leaving the MLP unrepaired collapses the cell. `attn_only` is therefore not a fair control across families, since it isolates a different noise source on each.

Across both pairs, `gate_up` keeps the most of the FT delta among the MLP-side masks (retention 0.78 and 0.73, against 0.63 and 0.54 for `mlp_only` and 0.31 and 0.19 for `ALL`) while capturing the top-two noise classes from App. L.1. It does not reach the Combined of `mlp_only` on Qwen or of `attn_only` on Llama. The best mask differs by pair, so no single mask is adopted, and the headline method keeps `ALL`, which repairs every matrix, as the untuned default.

L.3 Comparison with element-wise merging baselines

We compare DG-Hard against three element-wise merging baselines on two pairs, Qwen3.5-4B + Winogrande (severe FT) and Llama-3.2-3B + MedQA (mild FT). WiSE-FT and Task Arithmetic (Wortsman et al., 2022; Ilharco et al., 2023) are mathematically identical for a single (base, FT) pair, and we report them jointly. TIES (Yadav et al., 2023) keeps the top k entries of $|\Delta|$, sign-elects, and averages. DARE (Yu et al., 2024) randomly drops entries with probability p and rescales surviving entries by $1/(1-p)$. We use "DARE" to denote the canonical rescaled form (`rescale=True`) and "DARE-Linear" to denote the variant without rescaling (`rescale=False`).

Two operating points. Part A uses each baseline at its published default (WiSE-FT/Task-Arith α=0.5; TIES `keep_ratio`= 0.2 with λ=1; DARE `drop_prob`= 0.5 with `rescale=True`; DG-Hard at its parameter-free RMT-optimal threshold). TIES at its published keep ratio of 0.2 leaves the delta almost unchanged (r=1.000). Its canonical setting is therefore close to the fine-tuned model itself, and its scores should be read that way.

Matched rollback. Part B moves every method to the same Frobenius rollback r=0.5, which isolates shape effects from magnitude effects. Each method's knob is bisected on the cached singular values until the predicted retention reaches 0.5: α for WiSE-FT, the keep ratio for TIES, the drop probability for DARE-Linear, and for DG-Hard a scale s on the threshold, $\tau = s\,\tau^*$, with $s < 1$ keeping more of the spectrum than the closed-form cut. The r printed in each table is the retention realized on the repaired weights, which lands below the target for DARE-Linear (0.35) and for DG-Hard on Llama (0.43).

Realized retention. For DARE in Part B we use the linear no-rescale form (DARE-Linear), since DARE with `rescale=True` at p=0.5 has expected Frobenius ratio $\sqrt{1/(1-p)} \approx 1.41$, which is

incompatible with the matched r=0.5 target. The Part B DG-Hard row is therefore a scaled threshold used only for this matched comparison, and the parameter-free method is the canonical row.

What Part B tests. Part B asks one question: at the same amount of rollback, does reverting the singular directions that the threshold selects repair forgetting better than uniform interpolation or coordinate-wise merging edits? It does not bear on the closed-form cutoff of Part A. The scaled row is a tuned operating point, and whether the closed-form cut needs tuning is the question the spectral controls answer (§4.4, App. F). This matched-rollback comparison is the central methodological control.

Evaluation run. The ablations were evaluated in a separate run on five held-out benchmarks (ARC-C, GSM8K, HellaSwag, IFEval, MMLU), so their Base and FT scores and their damaged sets differ slightly from those of the headline run.

Table 28: Merging baselines on Qwen3.5-4B + Winogrande (severe FT damage). $D = \{\text{arc_challenge}, \text{gsm8k}\}$, $I = \{\text{hellaswag}, \text{ifeval}, \text{mmlu}\}$, $U = \emptyset$. Combined uses the protocol harmonic mean over the partition. **Bold** marks the best value per column within each part. DARE's (Yu et al., 2024) expected Frobenius ratio is $\sqrt{1/(1-p)} \approx 1.41$ at p=0.5. The canonical run averages three mask seeds, whose expected ratio is $\sqrt{1+1/3} \approx 1.155$. A Combined of 0 indicates a sub-statistic floored at 0 (e.g., % preserved non-positive on the improved partition), which collapses the harmonic mean.

Method	r	GSM8K	ARC-C	HSwag	IFEval	MMLU	task_win	Combined
Part A: canonical operating points								
WiSE-FT / Task-Arith α=0.5	0.500	**91.1**	**81.7**	60.9	35.5	72.6	85.7	69.1
TIES canonical	1.000	57.3	62.6	**87.1**	41.4	**75.8**	**89.8**	0.0
DARE canonical	1.155	56.3	65.8	87.0	**43.2**	75.3	89.6	10.3
DG-Hard canonical	0.314	87.7	73.9	85.0	36.1	75.7	89.4	**86.2**
Part B: matched Frobenius rollback r=0.5								
WiSE-FT / Task-Arith r=0.5	0.500	91.1	81.7	60.9	35.5	72.6	85.7	69.1
TIES r=0.5	0.504	**92.7**	**87.4**	47.3	35.2	65.0	68.0	7.5
DARE-Linear r=0.5	0.354	92.4	87.1	47.3	35.4	64.8	79.7	7.4
DG-Hard r=0.5	0.500	90.3	75.3	**87.1**	**38.3**	**79.8**	**89.3**	**92.0**

Table 29: Merging baselines on Llama-3.2-3B + MedQA (mild FT damage). $D = \{\text{gsm8k}\}$, $I = \{\text{hellaswag}, \text{mmlu}\}$, $U = \{\text{arc_challenge}, \text{ifeval}\}$. **Bold** marks the best value per column within each part. DARE's (Yu et al., 2024) expected Frobenius ratio is $\sqrt{1/(1-p)} \approx 1.41$ at p=0.5. The canonical run averages three mask seeds, whose expected ratio is $\sqrt{1+1/3} \approx 1.155$. A Combined of 0 indicates a sub-statistic floored at 0 (e.g., % preserved non-positive on the improved partition), which collapses the harmonic mean.

Method	r	GSM8K	ARC-C	HSwag	IFEval	MMLU	task_med	Combined
Part A: canonical operating points								
WiSE-FT / Task-Arith α=0.5	0.501	71.1	**74.6**	67.5	**76.9**	38.1	**58.7**	65.2
TIES canonical	1.000	67.2	74.5	70.1	75.2	**55.7**	58.1	5.4
DARE canonical	1.155	66.8	74.4	**70.2**	76.0	55.3	57.7	0.0
DG-Hard canonical	0.186	**73.8**	**74.6**	67.1	75.9	38.7	57.7	**71.8**
Part B: matched Frobenius rollback r=0.5								
WiSE-FT / Task-Arith r=0.5	0.501	71.1	74.6	67.5	76.9	38.1	**58.7**	65.2
TIES r=0.5	0.503	70.4	74.5	64.8	**78.3**	34.0	57.6	37.1
DARE-Linear r=0.5	0.354	70.1	**74.8**	65.3	78.0	31.3	58.0	21.9
DG-Hard r=0.5	0.427	**72.8**	74.1	**68.8**	75.9	**44.7**	57.4	**86.9**

Canonical operating points. On Qwen + Winogrande, DG-Hard wins on Combined (86.2 vs. 69.1 for WiSE-FT/Task-Arith) and improves the damaged benches relative to TIES/DARE (GSM8K 87.7 vs. 57.3/56.3; ARC-C 73.9 vs. 62.6/65.8). On Llama + MedQA, TIES and DARE score much higher on MMLU than DG-Hard (55.7/55.3 vs. 38.7), yet their Combined collapses to 5.4 and 0.0 respectively. The r column explains this. At canonical settings, TIES and DARE sit at $r{\approx}1$. They therefore do not shrink the FT delta and heal a negligible fraction of the FT-induced damage on GSM8K. The partition-conditional Combined headline reveals this gap, which an unweighted per-benchmark mean would miss. WiSE-FT and Task Arithmetic produce identical outputs for a single (base, FT) pair, so we report a single row and credit both citations.

Matched Frobenius rollback $r{=}0.5$. With magnitude held fixed, DG-Hard wins both pairs on Combined ($+22.9$ pp on Qwen + Winogrande, 92.0 vs. 69.1; $+21.7$ pp on Llama + MedQA, 86.9 vs. 65.2). The per-benchmark breakdown shows why. TIES at $r{=}0.5$ collapses HellaSwag (47.3 vs. 87.1 under DG-Hard). DARE-Linear at its realized rollback also collapses HellaSwag (47.3) and loses MMLU and on-task accuracy relative to DG-Hard. DG-Hard is the only method that reaches a comparable rollback without driving any of the five benchmarks to a degenerate value. TIES and DARE-Linear sparsify in the standard parameter basis, while DG-Hard sparsifies in the singular-vector basis. At a comparable Frobenius rollback, spectral sparsity outperforms coordinate sparsity.

M WHAT HAPPENS WHEN THE SIGNAL IS DIFFUSE?

DG-Hard keeps every singular direction of Δ above the Gavish-Donoho cutoff τ^*, however many there are. A broad task signal is therefore retained as long as it stands above the cutoff. The question is what happens to a task signal that does not, one spread thinly across many directions with singular values inside the noise bulk. This section collects the evidence in three parts: what the real fine-tuning deltas look like, what happens on synthetic matrices with a known signal as it approaches and crosses the noise floor, and what two controls that keep most of the spectrum recover on the real checkpoints.

Real deltas are spiky. On all 14 checkpoints the retained component is a small set of spikes. The cutoff keeps about 3% of the rank and 15.5% of the spectral energy per matrix (range 10.3 to 21.9% across checkpoints; Tab. 3), and on six checkpoints the discarded bulk matches the Marchenko-Pastur density on 85% of the matrices (Tab. 2, App. A). The two logical-reasoning tasks, ReClor and StrategyQA, keep 1.7 to 2.9% of the rank, as narrow as the other tasks. Prior reports of rank-deficient fine-tuning updates and low intrinsic dimension (Hu et al., 2022; Aghajanyan et al., 2021) and of a random-matrix bulk in trained weights (Thamm et al., 2022; Staats et al., 2023) describe the same picture. Nothing in the 14 measured deltas is diffuse. The 14 deltas come from two models, Qwen3.5-4B and Llama-3.2-3B, each fine-tuned on seven tasks.

Synthetic recovery. Each synthetic update is $\Delta = S + E$, where $S = U_s \operatorname{diag}(\mathbf{s})\, V_s^\top$ has exact rank r with orthonormal factors from a QR factorization of Gaussian matrices, and E is a full-rank matrix of IID standard Gaussian entries. We apply DG-Hard to Δ and report the relative reconstruction error $\|\hat{S} - S\|_F / \|S\|_F$ of the retained component against the true signal, averaged over five seeds. Tab. 30 holds the signal about 3.3 times above the noise edge and varies its breadth and the matrix size. Tab. 31 fixes rank 100 on a 512×512 matrix and varies the separation from the floor.

Breadth and size do not change the outcome. A rank-100 signal is recovered as accurately as a rank-20 one. Separation from the floor changes it, and this is the one regime in which the method fails. The error degrades gracefully as the signal approaches the floor. A signal whose spectrum decays through the floor is recovered on its above-floor part, and a signal buried below the floor is discarded entirely. A component inside the noise bulk is statistically indistinguishable from the noise, so no method that works from Δ alone can separate it, whatever its rank rule. This failure is synthetic. The real fine-tuning deltas sit in the regime of the first row of Tab. 30, where DG-Hard retains about 3% of the rank on average.

Spectrum-retaining controls on the real checkpoints. Tab. 32 shows that fixed energy-percentage at 90% and effective-rank, which keep most of the spectral energy, score far below DG-Hard on every cohort. Keeping the bulk therefore does not recover what DG-Hard recovers. Per benchmark,

Table 30: Synthetic recovery at a fixed margin above the noise floor, varying signal breadth and matrix size. Every signal direction is retained and the error stays near 0.21 throughout. The residual error comes from noise within the kept directions.

Synthetic signal	Rank kept	Rel. error
rank-20, 512×512 (matches real deltas)	3.9%	0.216
rank-100, 512×512 (broad)	19.5%	0.205
rank-8, 256×256	3.1%	0.214
rank-30, 384×384	7.8%	0.213
rank-24, 768×512 (non-square)	4.7%	0.215
rank-40, 1024×1024	3.9%	0.215

Table 31: Synthetic recovery at fixed rank 100 on a 512×512 matrix, varying the separation of the signal from the noise floor.

Signal relative to the noise floor	Directions kept	Rel. error
well above the floor ($\approx 3.3\times$)	$100/100$	0.21
near the floor ($\approx 1.8\times$)	$100/100$	0.40
diffuse, decaying through the floor	$71/100$	0.44
below the floor (buried)	$0/100$	1.00

DG-Hard heals 82.7% of the fine-tuning drop on GSM8K over its 8 damaged cells, 118.6% on MATH-500 over its 6, 79.6% on IFEval (5), 111.5% on TriviaQA (4), 79.4% on TruthfulQA (1), 33.5% on MNLI (4), and 27.9% on ARC-Challenge (2).

Table 32: Combined score per cohort for DG-Hard and the two controls that retain most of the spectrum. **Bold** marks the column maximum.

Method	Overall	Knowledge	Cognition	Qwen	Llama
DG-Hard	**86.1**	**84.0**	**87.5**	**86.4**	**78.8**
Fixed energy-percentage 90%	43.1	19.3	46.8	31.0	69.0
Effective-rank	32.1	2.5	36.4	18.1	62.5

N Positioning against data-based repair and unlearning

Tab. 33 compares DG-Hard with representative data-based methods along the axes that determine who can run the repair. Gradient ascent, NPO, and RSS each require the relevant data, gradient computation, and an optimization run, so they operate in the training-compute regime. Task-arithmetic negation is the closest in spirit, since its edit $\theta - \lambda\tau$ is a closed-form weight operation. Negating the given update itself is uniform interpolation, our WiSE-FT / Task Arithmetic baseline in Tab. 1, which is pair-only and data-free but leaves λ to be chosen. The row in Tab. 33 refers to negating a vector for the behavior to be removed, which requires a separate fine-tuning run on that behavior. DG-Hard targets broad capability and safety recovery from the two checkpoints alone. It is not a substitute for per-example trust repair when the failure traces to identifiable training examples and the data and compute to act on them are at hand. That truncated SVD is a standard tool does not make its use here incremental. The estimation problem is new, one update with an unknown noise scale and no data, and the answer to that problem, a calibrated per-matrix cutoff, is what the standard tool did not provide.

Table 33: Positioning against data-based repair and unlearning methods. "Pair alone" means the method repairs a given (base, FT) pair with no further artifacts.

Method	Data-free	Gradient-free	Closed-form	Pair alone	Target
DG-Hard (ours)	Yes	Yes	Yes	Yes	Capability and safety recovery
Gradient ascent (Jang et al., 2023)	No	No	No	No	Unlearn a forget set
NPO (Zhang et al., 2024b)	No	No	No	No	Unlearn a forget set
RSS (Shubham et al., 2026)	No	No	No	No	Trustworthiness (bias, ethics, truth)
Negation of a separately trained vector (Ilharco et al., 2023)	No	Edit only	Yes (the edit)	No	Model editing and merging