DIAGNOSING RECOVERABILITY: SUPPRESSION-LIKE VS. OVERWRITE-LIKE CAPABILITY LOSS IN LLM POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training improves task performance but can degrade general capabilities such as instruction following, factuality, and safety. The same accuracy drop can conceal two different states: a capability may remain accessible but difficult to elicit, or it may require substantial relearning. We call these *suppression-like* and *overwrite-like* forgetting. We introduce TRIAGE, a diagnostic and repair framework that combines recovery dynamics, hidden-state readability, and lightweight-repair operability into a continuous Forgetting Typing Score (FTS). FTS achieves 0.921 AUROC under leave-one-injection-type-out evaluation and transfers to new suppression mechanisms, a second model family, and a second training objective. At matched mathematical-reasoning gains, RLVR leaves more accessible losses than SFT, with FTS-weighted suppression indices of 74.2% and 38.7%. A controlled 2×2 study reproduces 79% of this gap within supervised training by combining on-policy data with proximity-to-base regularization; realized model–base distance strongly predicts forgetting type across supervised and RL runs. FTS-guided repair turns this distinction into a practical advantage, reaching a common $\approx 90\%$ recovery target at 1.7–$2.4\times$ lower cost than uniform replay. The type of forgetting thus connects how post-training changes a model to the intervention that restores its capabilities.

1 INTRODUCTION

Post-training gives LLMs new competence—SFT instills domain skills and RLVR drives reasoning gains (Shao et al., 2024; Guo et al., 2025)—but can degrade instruction following, factuality, safety, and multilingual competence (Luo et al., 2023; Qi et al., 2025; Ouyang et al., 2022). These losses are usually summarized by endpoint accuracy and treated with a common remedy (Rolnick et al., 2019; Wortsman et al., 2022; Ilharco et al., 2023). Yet equal drops can behave very differently: one model may recover after a few examples, while another remains impaired after extensive replay. The former is consistent with retained access, whereas the latter suggests greater disruption and relearning; forgetting should therefore be evaluated by both loss magnitude and recovery cost under new-task retention. No single signal fully resolves this distinction: recovery speed depends on task difficulty, hidden-state readability need not imply behavioral access, and lightweight operability depends on the available intervention family.

We introduce **TRIAGE** (**T**yping and **R**epair via **I**ntegrated **A**ssessment of **G**enuine and **E**phemeral forgetting), a framework that makes this distinction measurable. TRIAGE combines recovery dynamics, hidden-state readability, and lightweight operability into a continuous **Forgetting Typing Score (FTS)**. Under the fixed diagnostic and intervention protocol, high scores indicate suppression-like losses amenable to lightweight repair; low scores indicate overwrite-like losses requiring substantial replay. The score then guides the repair budget. This framing asks whether SFT and RLVR differ only in loss magnitude or also in recoverability; at matched task gains, we find a systematic accessibility gap.

Contributions. (1) **A predictive diagnosis of forgetting type**: FTS combines three repair-relevant signals and generalizes across mechanisms, model families, and objectives (§4.1). (2) **An explanation of the SFT–RLVR gap**: matched-gain comparisons reveal a 35.5-point separation in

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.

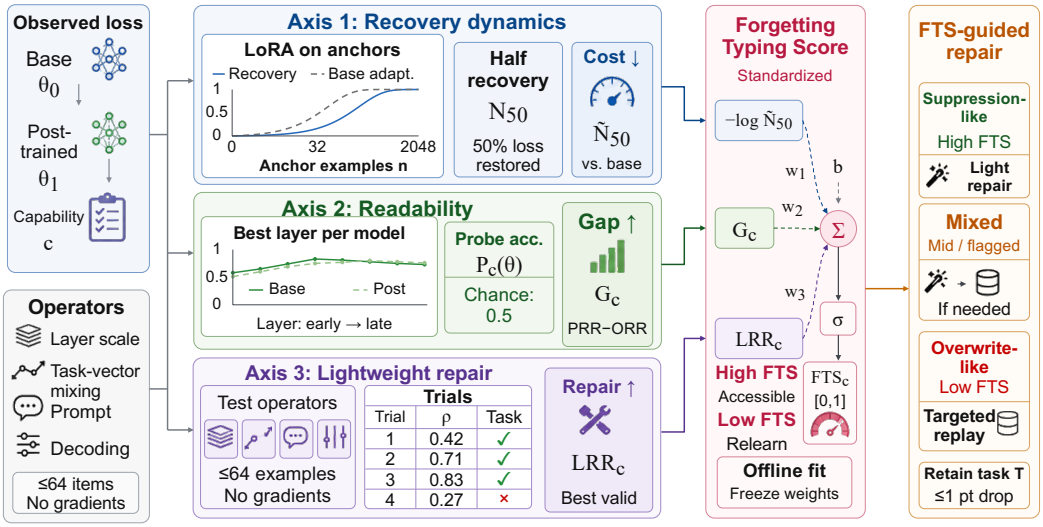

Figure 1: TRIAGE overview. Three complementary axes yield a frozen FTS under a fixed intervention family and guide repair. Curves and trial scores are schematic. Axis 1 uses $\log(1+n)$ spacing; its dashed curve shows ceiling-normalized base adaptation.

suppression-like loss, and a controlled factorial identifies on-policy data and proximity to the base model as the main explanatory factors (§4.2, §4.3). (3) **Diagnosis-driven repair**: type-aware routing directs lightweight interventions toward accessible losses and concentrates replay on difficult ones, reducing cost by 1.7–2.4× at a common $\approx 90\%$ recovery target (§4.4).

2 Related Work

Forgetting phenomena and explanations. Catastrophic forgetting is well documented for LLMs (McCloskey & Cohen, 1989; French, 1999; Goodfellow et al., 2014; Luo et al., 2023; Kotha et al., 2024). Recent work shows that apparent knowledge loss can reflect task-alignment or readout drift (Zheng et al., 2025), while analyses such as RL's Razor (Shenfeld et al., 2026) and on-policy retention (Chen et al., 2026) explain why some post-training objectives preserve base behavior. Complementarily, Harmon et al. (2026) decompose aggregate accuracy changes into sample-wise forgetting ($1 \to 0$ transitions) and backward transfer ($0 \to 1$ transitions), revealing changes hidden by task averages. We ask a different question conditional on an observed loss: whether the failed capability remains cheaply recoverable. We compare their item-level metric with FTS as a predictor of recovery type.

Reversibility and hidden-state evidence. Work on unlearning distinguishes reversible and irreversible effects using representation similarity (Xu et al., 2026; Kornblith et al., 2019). Separately, models can encode information that they fail to express (Orgad et al., 2025; Qi et al., 2025; Lin et al., 2024; Yun et al., 2025). TRIAGE integrates probing (Alain & Bengio, 2017; Belinkov, 2022), recovery dynamics, and direct operability into one calibrated score. Control tasks (Hewitt & Liang, 2019) and causal patching test whether probe readability corresponds to recoverable behavior.

Repair techniques. LiNeS (Wang et al., 2025), task-vector merging (Ilharco et al., 2023; Wortsman et al., 2022), and replay (Rolnick et al., 2019) provide complementary repair mechanisms. TRIAGE connects these interventions through diagnosis: it identifies capabilities accessible to lightweight repair and directs replay toward losses that require relearning.

3 The TRIAGE Framework

TRIAGE measures *accessibility* to guide repair (Figure 1): how quickly behavior returns, how much task-relevant information remains readable, and how effectively lightweight interventions restore

performance. A one-time offline calibration freezes the score specification on controlled reference states; per-checkpoint diagnosis then measures the three axes, scores each degraded capability, and routes repair (Appendix B.6).

3.1 From Capability Loss to Recovery

Let θ_0 be a base model and $\theta_1 = \mathrm{Post}(\theta_0; T)$ its post-trained version. For capability c, evaluator $A_c(\theta) \in [0, 1]$ measures performance and $\Delta_c = A_c(\theta_0) - A_c(\theta_1)$ measures the loss. The unit of diagnosis is a model–checkpoint–capability tuple with $\Delta_c \geq 2$ points in the three-seed mean. We call readily accessible losses *suppression-like* and losses requiring substantial relearning *overwrite-like*. FTS locates each loss along this recoverability continuum under a fixed intervention family.

Repair options. The lightweight family $\mathcal{R}^{\mathrm{v1.0}}_{\mathrm{light}}$ combines parameter-space correction (per-layer task-vector scaling and global interpolation) with inference-space correction (native prompt/template restoration and decoding interventions) (Wang et al., 2025; Ilharco et al., 2023; Wortsman et al., 2022; Lin et al., 2024). These operators are gradient-free and use at most $B_{\mathrm{light}} = 64$ capability-validation examples for selection. The heavyweight alternative $\mathcal{R}_{\mathrm{heavy}}$ is replay fine-tuning on capability anchors: examples that exercise c and are disjoint from evaluation. Both aim to restore c while preserving the gain on the newly trained task T. For an intervened model θ', we measure

$$\rho_c(\theta') = \frac{A_c(\theta') - A_c(\theta_1)}{\Delta_c}, \qquad \text{s.t. } A_T(\theta_1) - A_T(\theta') \leq \epsilon_T, \tag{1}$$

where $\rho_c = 1$ denotes full recovery to the base model's performance and $\epsilon_T = 1$ point is the new-task retention budget. A lightweight operator reaches its repair target at $\rho_c \geq \eta = 0.8$. Appendix B.1 specifies the operator grids.

3.2 Three Complementary Measurements of Accessibility

Axis 1: how quickly does behavior return? We fine-tune θ_1 on capability anchors with LoRA and measure recovery as the number of examples increases. The calibration grid is $n \in \{0, 8, 32, 128, 512, 2048\}$; fitting $\rho_c(n) = \rho_\infty(1 - e^{-n/\kappa})$ summarizes the recovery ceiling ρ_∞ and learning scale κ. Our primary statistic, N_{50}, is the number of examples needed to restore half of the lost performance. Fast recovery indicates a retained head start; slow recovery indicates greater relearning demand (Figure 3). Each budget point is an independent LoRA fine-tune of θ_1 on a prefix-nested subset of the anchor pool; curves whose fitted ceiling falls below half recovery enter the score right-censored at $4{,}096$ examples (Appendix B.3).

Capabilities differ in adaptation speed. We therefore normalize N_{50} by the clean base model's adaptation half-time N_{50}^{ref}: the number of examples needed to attain half of its fitted asymptotic improvement on the same capability anchors under the same LoRA protocol:

$$\tilde{N}_{50} = N_{50}/N_{50}^{\mathrm{ref}}. \tag{2}$$

A smaller ratio indicates faster recovery relative to the capability's reference adaptation timescale. Reference curves are shared across checkpoints from the same base model. For repair deployment, a compressed grid $\{0, 6, 24, 96, 384\}$ reduces measurement cost; Appendix B.3 gives the fitting procedure and Appendix E the training and cost settings.

Axis 2: does the model still represent the relevant information? We construct a balanced binary task for each capability and train a linear probe on the last-token residual-stream representation. For knowledge, the input is a question paired with an answer option, and the label indicates whether the option is correct. For instruction following, the input is an instruction–response pair, and the label indicates whether the response satisfies the constraint. These tasks measure the readability of information needed for the original behavior.

Let $P_c(\theta)$ be the probe's mean accuracy under nested five-fold question-level cross-validation, with layer and regularization selected separately for each model. Independent layer selection asks whether relevant information remains linearly available anywhere in the model, without requiring it to stay at the same depth. Chance normalization gives $\tilde{P}_c(\theta) = (P_c(\theta) - P_{\mathrm{chance}})/(1 - P_{\mathrm{chance}})$. We compare internal and behavioral retention through $\mathrm{PRR} = \tilde{P}_c(\theta_1)/\tilde{P}_c(\theta_0)$ and $\mathrm{ORR} =$

$A_c(\theta_1)/A_c(\theta_0)$, and use their difference, $G_c = \mathrm{PRR} - \mathrm{ORR}$, as the *readout gap*. A large positive gap identifies information that remains readable despite weaker outputs. Outside the calibration family, probes transfer through a per-layer linear correspondence map or are refit locally at negligible cost (Appendix B.4.1). Appendix B.4 details the probe construction; Appendix D.3 connects readability to behavioral recovery.

Axis 3: which lightweight repair can expose the capability? We evaluate the configurations of $\mathcal{R}^{\mathrm{v1.0}}_{\mathrm{light}}$ on the capability-validation slice, retain those satisfying the new-task constraint, and record the best recovery:

$$\mathrm{LRR}_c = \max_{r \in \mathcal{R}^{\mathrm{v1.0}}_{\mathrm{light}}} \rho_c(r(\theta_1)) \quad \text{s.t. } A_T(\theta_1) - A_T(r(\theta_1)) \le \epsilon_T. \tag{3}$$

This *lightweight repair rate* measures recovery without gradient updates. We cache every candidate's validation score and the maximizing configuration r_c for routing. Whereas LRR_c measures recovery by available operators, dynamics quantify remaining learning demand and probes assess retained information. Agreement supports confident typing; disagreement routes the capability to the mixed tier for continuous repair-budget allocation.

3.3 Calibrating the Forgetting Typing Score

Reference construction. Calibration connects the measurements to two reference recovery states. We measure p_c, recovery from transplanting a width-4 window of clean θ_0 activations, and u_c, recovery after replay with 4,096 examples for two epochs. Patching asks whether retained circuitry can be re-exposed without relearning; the replay ceiling asks how much relearning the loss demands. The suppression reference combines $p_c \ge 0.70$ with $u_c \ge 0.50$, the overwrite reference values below both, yielding 52 references of each type on which the fusion weights and standardization statistics are fitted once and frozen as v1.0. The reference recipes, mixed cases, and five boundary cases within this set are described in Appendices C and D.9.

Fusion. Logistic regression combines the three standardized measurements:

$$\mathrm{FTS}_c = \sigma\big(w_1\, z(-\log \tilde{N}_{50}) + w_2\, z(G_c) + w_3\, z(\mathrm{LRR}_c) + b\big) \in [0,1], \tag{4}$$

where σ is the logistic sigmoid and z uses calibration-set statistics. Faster recovery, a larger readout gap, and stronger lightweight repair all raise FTS. After calibration, the weights and standardization remain fixed across checkpoints, model families, and training objectives (Appendix B.2). Scores above 0.7 are suppression-dominant, scores below 0.3 are overwrite-dominant, and the intervening band captures mixed losses. Reliability flags also assign cases to the mixed tier (Appendix D.9).

3.4 Turning Diagnosis into Repair

FTS guides a two-stage repair policy.

Stage 1: restore access. Suppression-dominant and mixed capabilities receive lightweight repair. Global parameter operators are chosen jointly from the cached candidate grid, maximizing loss-weighted validation recovery across the receiving capabilities. Prompt and decoding configurations are applied per capability. A held-out check retains repairs reaching η for suppression-dominant capabilities or positive recovery for mixed capabilities, subject to new-task retention.

Stage 2: concentrate replay. Overwrite-dominant and unresolved mixed losses receive anchor examples in proportion to $(1 - \mathrm{FTS}_c)\Delta_c$, within the total replay budget B. Replay starts from the Stage-1 model, mixes in 15% new-task data, and uses KL regularization to that model. A final held-out check evaluates recovery and retention; one additional allocation round is available if the target remains unmet and budget remains. The router returns a single repaired checkpoint and the per-capability inference-time configurations selected in Stage 1 (Algorithm 1; Appendix B.6).

4 Experiments: Diagnosis, Mechanism, and Repair

We first establish that FTS identifies recovery states, then use it to explain the SFT–RLVR difference and guide repair (Figure 2). **Setup.** The main study post-trains Qwen3-8B on mathematical reasoning: SFT uses NuminaMath/OpenR1-Math traces (Li et al., 2024), and RLVR uses GRPO (Shao

et al., 2024) on DeepScaleR prompts (Luo et al., 2025). We evaluate seven retained capabilities covering knowledge, instruction following, chat, factuality, safety, and multilingual reasoning. Comparisons stop at matched task gain and use three seeds unless noted. Transfer covers Qwen3-1.7B/4B/14B (Yang et al., 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and instruction-following post-training. Appendix E gives the protocols; Appendix A lists the evaluation pools.

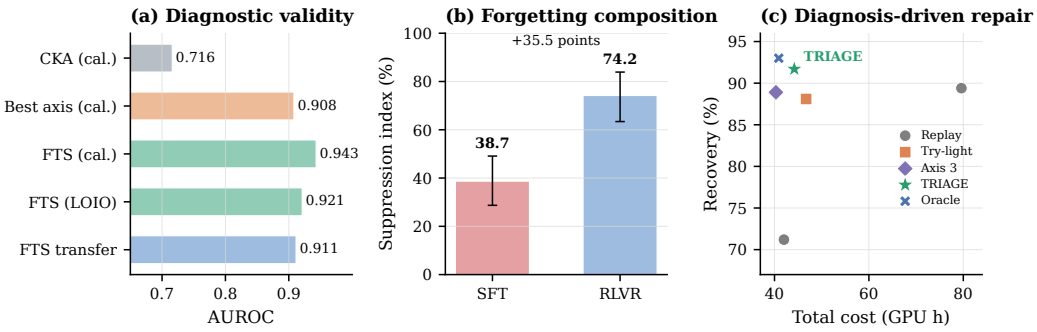

Figure 2: Empirical summary. **(a)** FTS outperforms individual axes and CKA on the calibration set, and remains strong under LOIO and transfer. **(b)** RLVR losses are more suppression-like at matched gain. **(c)** TRIAGE improves the repair cost–recovery trade-off. LOIO leaves out injection types; transfer averages the three AUROC evaluations in Appendix Table 9; bars in (b) show 95% CIs.

4.1 FTS Identifies Distinct Recovery States

Separation on controlled references. Five injection recipes span format/readout drift, alignment drift, top-layer shift, counterfactual updates, and gradient-ascent unlearning (Yun et al., 2025; Lin et al., 2024; Zheng et al., 2025; Qi et al., 2025; Meng et al., 2022; Jang et al., 2023). The resulting reference states have visibly different recovery dynamics: suppression-like losses recover rapidly, while overwrite-like losses require more anchor data, with an $11.5\times$ difference in median pair-normalized half-recovery cost (Figure 3).

Generalization by construction. Reference labels derive solely from patching and replay-ceiling measurements, transfer pools never enter fitting, and the v1.0 specification remains frozen across every subsequent evaluation. Withheld mechanisms (LOIO), continuous mixtures, construction-independent perturbations, and non-invertible suppressions then test the score on recovery states it has never seen (Appendices C–D).

Under leave-one-injection-type-out evaluation, FTS achieves AUROC 0.921 ± 0.018, versus 0.719 ± 0.053 for sample-wise forgetting (Harmon et al., 2026). After fitting on the full 104-reference calibration suite, FTS reaches AUROC 0.943 $[0.914, 0.968]$ on that same set, compared with 0.908 for the strongest individual axis (DeLong $p = 0.008$), 0.716 for CKA, and 0.760 for sample-wise forgetting. FTS thus captures recovery structure that endpoint transitions and representation similarity alone leave unresolved (Appendix Table 9).

Transfer to new recovery states. The fixed score types $16/18$ suppressions that admit no operator inverse, and reaches AUROC 0.917 on Llama-3.1-8B-Instruct and 0.905 on the instruction-following objective. In $13/18$ non-invertible cases the successful operator is not the perturbation's structural counterpart: FTS tracks recoverability itself, not a specific reversal path. On real post-training checkpoints, 36 of 39 externally verified instances agree with the reference verdict, and the models' own outputs tell the same story: the surface-form share of newly failed items correlates with FTS at $\rho = 0.81$, with high-FTS losses changing how answers are expressed and low-FTS losses changing what they say (Appendices D.4 and C.5).

Connecting readability to behavior. Probe retention tracks causal patching recovery at $r = 0.74$ on calibration references and $r = 0.71$ on the high-rank real-checkpoint stratum: the information Axis 2 reads out is precisely the information that activation restoration returns to behavior. Instances whose damage spectrum makes this link least stable are routed to the mixed tier by the reliability

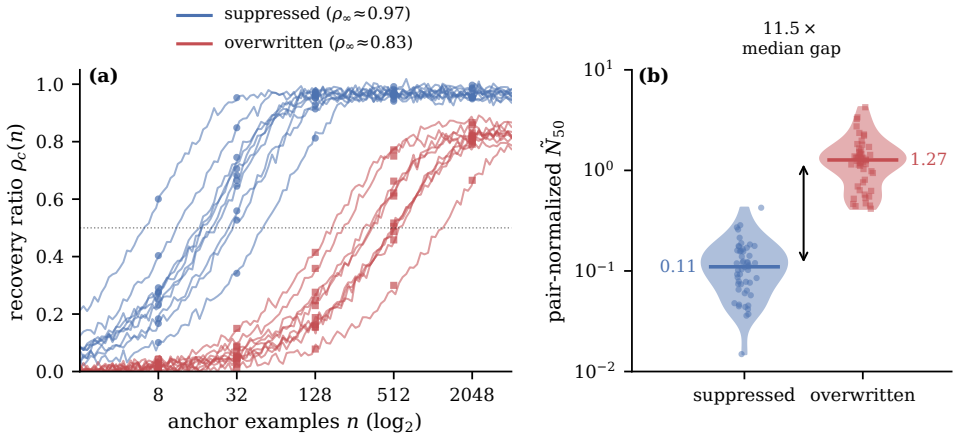

Figure 3: Recovery dynamics underlying Axis 1. **(a)** Suppression-like references recover rapidly and saturate at a higher ceiling, while overwrite-like references require substantially more anchor data. **(b)** Pair-normalized half-recovery cost separates the 104 reference instances by $11.5\times$ in median.

flags, so the continuous score stays informative exactly where a binary typing would overcommit (Appendix D.3).

Stability across measurement choices. Varying the reference-measurement protocol—patching window, recovery threshold, and replay budget—moves at most 9 of 104 reference assignments and keeps AUROC above 0.92 (Appendix D.6). Removing any single operator class from the lightweight family preserves the accessibility ranking at Kendall-$\tau \geq 0.93$; adding SVD projection changes 3.2% of type assignments (Appendix D.8). The score captures recoverability structure that is stable across these measurement and operator-family choices.

4.2 What Kind of Forgetting Do SFT and RLVR Produce?

At matched math gain (GSM8K +11.8/+12.3; MATH500 +9.6/+9.1), SFT and GRPO leave markedly different recovery states. We summarize these states with the FTS-weighted suppression index $\sum_c \mathrm{FTS}_c \Delta_c / \sum_c \Delta_c$: the mean FTS weighted by each capability's performance loss Δ_c, expressed as a percentage. Table 1 reports both loss magnitude and forgetting type for each capability.

Table 1: Forgetting anatomy at matched math gain (Qwen3-8B). Type from FTS: S = suppression-dominant, O = overwrite-dominant, M = mixed; $\Delta\mathrm{FTS} = \mathrm{FTS}^{\mathrm{RLVR}} - \mathrm{FTS}^{\mathrm{SFT}}$; mean ± s.d. over 3 seeds. The suppression index is FTS-weighted with capability-clustered bootstrap 95% CIs.

	SFT (matched gain)			**RLVR / GRPO (matched gain)**			
Capability	Δ_c	FTS	**T**	Δ_c	FTS	**T**	ΔFTS
MMLU (knowledge)	5.8±0.4	0.18	O	2.1±0.3	0.58	M	+0.40
MMLU-Pro	7.3±0.6	0.12	O	2.4±0.4	0.49	M	+0.37
IFEval (instr.)	8.9±0.7	0.44	M	3.1±0.5	0.87	S	+0.43
AlpacaEval 2 LC	9.6±1.1	0.58	M	2.8±0.6	0.92	S	+0.34
TruthfulQA	6.2±0.5	0.24	O	2.2±0.4	0.68	M	+0.44
Safety (AdvBench+XSTest)	7.8±0.9	0.71	S	2.6±0.5	0.93	S	+0.22
MGSM (multiling.)	4.7±0.6	0.22	O	2.0±0.3	0.59	M	+0.37
Mean degradation	7.2 points			2.5 points			
FTS-weighted suppression index	**38.7**% [28.7, 49.1]			**74.2**% [63.4, 83.9]			

RLVR leaves more accessible losses. Its suppression index is 74.2% [63.4, 83.9], compared with 38.7% [28.7, 49.1] for SFT, a 35.5-point gap (permutation $p = 0.004$). FTS is higher under RLVR

for all seven capabilities, with paired differences from $+0.22$ to $+0.44$ (exact sign test $p = 0.016$). Under SFT, knowledge, factuality, and multilingual competence are the most overwrite-like losses, while instruction-following, chat, and safety remain comparatively accessible.

Training changes accessibility as well as accuracy. Mean degradation under RLVR is smaller (2.5 versus 7.2 points), and an equal-severity comparison reveals a composition difference too: early SFT matched to GRPO degradation reaches a 49.3% suppression index, 24.9 points below GRPO. Along the trajectories, SFT falls from 59.2% to 38.7%, while GRPO stays near 74–78% (Appendix F.1).

Robustness of the ordering. The gap survives every alternative summary we examined: even assigning all mixed-band loss to the overwrite pole leaves 24.9 points, and propagated evaluator noise or a small-loss damping of the recovery denominator moves the indices by under two points (Appendix J).

4.3 What Produces the SFT–RLVR Difference?

We isolate two features that distinguish the training pipelines: whose trajectories provide the targets and how strongly training remains near the base model. A matched 2×2 factorial within supervised training crosses teacher-written versus model-generated traces with and without KL-to-base regularization, holding prompts, verifier, and token budgets approximately fixed. Additional implementations are reported in Appendix G.

Table 2: Factorial decomposition of the SFT–RLVR gap (Qwen3-8B, matched math gain, fixed FTS v1.0, 3 seeds). $\|\delta\|$ is relative to the main SFT arm, and KL is measured to θ_0. † marks the matched 2×2; ext.* denotes verifier-filtered teacher traces on the same prompts. Full sweeps appear in Table 23.

Arm	Data	Prox. reg.	Mean Δ_c	$\|\delta\|$	Suppr. index	KL
SFT (main)	external	—	7.2	1.00	38.7% [28.7, 49.1]	0.94
† SFT, matched prompts/tokens	ext.*	—	6.8	0.92	40.5% [30.2, 51.0]	0.91
† SFT, matched + KL	ext.*	KL β=0.01	5.3	0.70	53.1% [42.4, 63.5]	0.51
† SFT, self-generated	on-policy	—	5.1	0.69	58.1% [47.5, 68.2]	0.47
† **SFT, self-gen. + KL**	**on-policy**	**KL** β=0.01	3.2	0.54	**68.5**% [58.1, 78.0]	0.29
GRPO, $\beta = 0$	on-policy	—	3.2	0.51	66.9% [56.3, 76.8]	0.33
GRPO, $\beta = 0.01$ (main)	on-policy	KL β=0.01	2.5	0.38	74.2% [63.4, 83.9]	0.21

Two training choices reproduce most of the gap. Model-generated traces raise the suppression index by 16.5 points and proximity regularization by 11.5 points, with a small interaction (-2.2 points, $p = 0.34$). Together they move supervised training from 40.5% to 68.5%, reproducing 79% of the original SFT–RLVR gap (95% CI $[73, 83]$). The factors also reach different capabilities: self-generated traces produce larger FTS gains on knowledge and factuality, while KL regularization produces larger gains on chat and safety. For MMLU, self-generation raises FTS from 0.21 to 0.42, versus 0.33 with KL alone; for chat, KL raises it from 0.62 to 0.80, versus 0.69 with self-generation (Appendix G).

Realized proximity unifies the training arms. Across 13 matched-gain supervised and RL arms, measured KL to the base model predicts the suppression index with $R^2 = 0.96$; adding trace source raises this to 0.98. A held-out PPO arm is predicted at 69.7% and measured at 71.8%. The relation persists within the nine supervised arms ($R^2 = 0.936$), so pooling objectives is not its sole source (Appendix G). The pre-fusion measurements show the same pattern: normalized half-recovery cost rises with measured KL ($\rho = +0.63$) while the readout gap falls ($\rho = -0.57$; Appendix Table 16), so the proximity–accessibility link holds at the level of individual axes, not only in the fused score.

Training choices also change repair cost. Repairing the self-generated SFT+KL checkpoint to the common recovery target costs 43.4 GPU·h, compared with 71.3 for main SFT; replay's share of the total falls from 84% to 61%.

4.4 DIAGNOSIS ENABLES EFFICIENT REPAIR

We evaluate repair on SFT- and RLVR-degraded Qwen3-8B checkpoints over three seeds. Table 3 compares complete policies at similar cost and at a common $\approx 90\%$ recovery target, including diagnosis, intervention, replay, validation, and safeguards.

Table 3: Core closed-loop comparison on Qwen3-8B. Costs are GPU·h and include diagnosis, repair, validation, and safeguards; rows pool the SFT- and RLVR-degraded scenarios over three seeds. The complete itemization and additional ablations are in Appendix Table 25.

Method	Total cost	Recovery	Retention
Uniform replay (comparable cost)	42.0	71.2±2.6%	98.1%
Uniform replay ($\approx 90\%$ target)	79.6	89.4±1.8%	97.3%
No-diagnosis router (FTS $\equiv 0.5$)	44.1	78.2±3.0%	98.4%
Try-light-first, then replay	46.7	88.1±2.7%	98.5%
Axis-3-only router	40.3	88.9±2.4%	98.7%
TRIAGE (full)	**44.2**	**91.7±1.5%**	**99.1%**
Oracle router	40.9	93.0±1.2%	99.2%

Recovering capabilities at lower cost. At approximately 90% recovery, TRIAGE costs 17.1 versus 41.5 GPU·h on RLVR-degraded checkpoints and 71.3 versus 117.7 on SFT-degraded checkpoints: $2.4\times$ and $1.7\times$ savings, respectively. These totals include 5.9 GPU·h of diagnosis; charging reference curves afresh for a new base raises diagnosis to 9.8 GPU·h, with savings still $2.0\times$ and $1.6\times$ (Appendix E.3). In the pooled comparison, it also retains more of the new-task gain (99.1% versus 97.3%). The larger saving on RLVR checkpoints follows directly from their more accessible loss profile.

What multi-axis diagnosis adds. Lightweight interventions lower recovery cost, and multi-axis diagnosis sharpens the allocation between lightweight repair and replay: try-light-first reaches 88.1% recovery and the Axis-3-only router 88.9%, against 91.7% for TRIAGE. The fusion margin is largest exactly where direct operator success is least decisive—4.7 points across model families and 6.8 on mixed-band losses—and cost-matched extrapolations preserve the method ordering (Appendix H.1.1).

Matching repair to recovery state. Table 4 exposes the complementary cost structure. Lightweight operators recover 88.4% of suppression-like losses but only 13.7% of overwrite-like losses. Replay also restores suppression-like losses, but costs $7.4\times$ more per recovered point and sacrifices 2.1 new-task points when the retention gate is released. Diagnosis therefore selects not simply a successful repair, but an economical one compatible with retaining the new task.

Table 4: Repair–type cross-application (Qwen3-8B, both damage scenarios pooled). Cost per recovered point is in GPU·h; Δtask is the new-task score change. Unlike the closed-loop policies in Table 3, this diagnostic experiment releases the retention gate to expose the cost of mismatched repair.

Diagnosed type	Repair applied	Recovery	Cost / point	Δtask
Suppression-like	light operators ($\mathcal{R}^{\mathrm{v1.0}}_{\mathrm{light}}$)	88.4±2.3%	0.05	-0.2
Suppression-like	large-budget replay	90.2±1.9%	0.37	-2.1
Overwrite-like	light operators ($\mathcal{R}^{\mathrm{v1.0}}_{\mathrm{light}}$)	13.7±3.6%	0.29	-0.3
Overwrite-like	targeted replay	84.9±2.5%	0.31	-0.8

The closed loop runs on diagnostic quality. Supplying oracle types to the same router yields 93.0% recovery, only 1.3 points above FTS: the router already extracts nearly all of the value available in the type information. Corrupting the diagnosis moves performance in proportion—permuting FTS drops recovery to 63.8%, and flipping 20% of assignments costs 11.8 recovery points and 38% extra cost—locating the closed-loop gain squarely in the diagnostic signal rather than in the routing machinery (Appendix H.2).

4.5 Transfer Across Models, Granularities, and Objectives

Across models. The fixed FTS specification transfers from Qwen3-1.7B to 14B and to Llama-3.1-8B-Instruct. The RLVR–SFT ordering holds on all five models; across the Qwen series the composition gap spans 29.3–36.3 points, and diagnosis-driven repair retains a 1.5–$2.0\times$ cost advantage. At both 4B and 14B, regularized on-policy SFT also remains within 6.9 points of GRPO in suppression index (Appendix F.1).

Across granularities. The same ordering appears across four MMLU macro-categories and ten representative disciplines, where category-level routing improves recovery by 2.1–2.6 points at fixed budget (Appendix F.2).

Across objectives. We also switch post-training from mathematical reasoning to verifiable instruction following, replacing instruction/chat in the retained capability set with mathematical reasoning. At matched IFEval gain, GRPO has a 66.3% suppression index versus 45.6% for SFT, a 20.7-point gap. Mathematical reasoning becomes one of the most overwrite-like losses under instruction SFT (FTS 0.22–0.26). The detailed loss profile therefore changes with the training target while the paradigm ordering persists. With frozen FTS weights and the same routing thresholds, the complete loop reaches 90.1% recovery with a $1.9\times$ cost advantage over replay and a $+2.9$-point fusion margin. The diagnosis transfers not only as a ranking, but as a repair decision (Appendices I.1 and I.2).

5 Discussion

Implications for post-training. Two runs with similar task gain or capability loss can leave the base model in states with sharply different repair costs. Our factorial reframes part of the categorical "RL versus supervised learning" contrast in terms of two controllable variables: whose trajectories supply the targets and how far optimization moves from the base. Within the tested regimes, realized model–base distance is more informative than the nominal objective or regularization coefficient.

Implications for maintenance. Uniform replay spends its largest budget on every loss; a type-aware policy reserves replay for losses that show little evidence of retained accessibility and uses lightweight interventions elsewhere. The benefit grows when a deployment maintains many checkpoints or capabilities, because diagnosis is shared across a portfolio of repair decisions. Category-level results further show that heterogeneous sub-capabilities can receive different budgets under the same rule.

Scope. Suppression-like and overwrite-like operationalize recoverability relative to the v1.0 reference protocol and intervention family; enlarging the operator family or the reference budget extends the score rather than invalidating it. The present evidence spans two model families, five scales, and two post-training objectives with frozen weights throughout, and the protocol extends directly to further architectures and training regimes. Near-threshold losses are routed to the mixed tier by design, where continuous replay allocation turns typing uncertainty into a proportional budget rather than a binary risk.

6 Conclusion

Post-training changes not only how much capability a model retains, but how accessible its remaining capabilities are. TRIAGE measures this second dimension, reveals the different recovery states left by SFT and RLVR, and links them to controllable training choices. The same diagnosis guides repair, restoring accessible losses with lightweight interventions and concentrating replay on relearning-intensive ones. Evaluating forgetting through recoverability therefore improves both our understanding of post-training and our ability to maintain model capabilities.

AI USE STATEMENT

We developed the theoretical framework, formulated its mathematical claims, designed the experimental methodology, implemented the method, and interpreted the results, using generative AI tools throughout as a discussion partner: each step proceeded through iterative exchanges in which we posed the questions, assessed the suggestions, and decided what to keep. We also used the same tools to draft and edit parts of the manuscript, including language polishing, length compression, and internal consistency checking, and for general coding assistance. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work preserves safety-relevant capabilities during post-training using only public benchmarks. Identifying suppressed alignment could inform unmasking, but all interventions here restore aligned behavior, and the diagnostic requires white-box access model owners already possess.

REPRODUCIBILITY STATEMENT

The exact objective is given in Sec. 3, and implementation details and hyperparameters are given in Appendices B and E. Prompt/template restoration and its search settings are described in Appendix B.1. Code for the proposed method and experimental pipeline is included in the supplementary material.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (Workshop)*, 2017.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. In *International Conference on Machine Learning (ICML)*, 2026.

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

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Jackson Harmon, Andreas Hochlehnert, Matthias Bethge, and Ameya Prabhu. Mapping post-training forgetting in language models at scale. In *International Conference on Learning Representations (ICLR)*, 2026.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Jian Hu, Xibin Wu, Wei Shen, Jason Klein Liu, Zilin Zhu, Weixun Wang, Songlin Jiang, Haoran Wang, Hao Chen, Bin Chen, Weikai Fang, Xianyu, Yu Cao, Haotian Xu, and Yiming Liu. OpenRLHF: An easy-to-use, scalable and high-performance RLHF framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 656–666. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-demos.48. URL https://aclanthology.org/2025.emnlp-demos.48/.

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

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. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805/.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Suhas Kotha, Jacob M. Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://iclr.cc/virtual/2024/poster/18492.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. NuminaMath. Hugging Face dataset release, 2024. URL https://github.com/project-numina/aimo-progress-prize.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://iclr.cc/virtual/2024/poster/17473.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. DeepScaleR: Surpassing O1-Preview with a 1.5B model by scaling RL. Notion Blog, 2025. URL https://pretty-radio-b75.notion.site/19681902c1468005bed8ca303013a4e2.

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. *arXiv preprint arXiv:2308.08747*, 2023.

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.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://iclr.cc/virtual/2025/poster/30060.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://iclr.cc/virtual/2025/poster/30893.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, 2019.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300v3.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=7HNRYT4V44.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, pp. 1279–1297. ACM, 2025. doi: 10.1145/3689031.3696075.

Ke Wang, Nikolaos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. LiNeS: Post-training layer scaling prevents forgetting and enhances model merging. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.17146v2.

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, 2022. URL https://arxiv.org/abs/2109.01903.

Xiaoyu Xu, Xiang Yue, Yang Liu, Qingqing Ye, Huadi Zheng, Peizhao Hu, Minxin Du, and Haibo Hu. Unlearning isn't deletion: Investigating reversibility of machine unlearning in LLMs. In *International Conference on Machine Learning (ICML)*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Longfei Yun, Chenyang An, Zilong Wang, Letian Peng, and Jingbo Shang. The price of format: Diversity collapse in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15454–15468. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.836. URL https://aclanthology.org/2025.findings-emnlp.836/.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious forgetting in continual learning of language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://iclr.cc/virtual/2025/poster/29593.

APPENDIX

APPENDIX GUIDE

The appendix separates implementation, evidence, and statistical analysis. Appendix A maps the evaluation pools. Appendices B, C, and D describe the diagnostic, reference construction, and validation; Appendix B.6 summarizes the end-to-end protocol from calibration to deployment. Appendix E specifies training and compute accounting. The remaining sections expand forgetting anatomy (Appendix F.1), the training-factor study (Appendix G), repair (Appendix H.1), and transfer (Appendices I.1 and I.2); Appendix J collects statistical procedures.

A EVALUATION SETS AND DATA SPLITS

Table 5 summarizes the calibration, transfer, and real-checkpoint evaluation pools and records which subsets enter fitting or external validation.

B DIAGNOSTIC IMPLEMENTATION

B.1 LIGHTWEIGHT OPERATOR FAMILY

Table 6 gives the v1.0 operator specification used throughout the paper. All operators are gradient-free at repair time; tuning uses at most $B_{\text{light}} = 64$ held-out validation examples per capability and zero backpropagation through the model.

Constraint handling: every tuning objective is ρ_c subject to $A_T(\theta_1) - A_T(\theta') \leq \epsilon_T = 1$ point, evaluated on a 128-example new-task dev slice; violating configurations are discarded before argmax.

Additional operator. SVD low-rank projection improves recovery by more than 0.15 over the v1.0 family on 6% of calibration instances, with higher seed variance on safety and chat. Appendix D.8 reports the ranking changes from adding it.

B.2 SCORE SPECIFICATION AND COMPARABILITY

FTS v1.0 consists of the operator family in Table 6, calibration-set standardization statistics, and logistic weights $(w_1, w_2, w_3, b) = (1.84, 0.96, 1.37, -0.21)$. All checkpoint, family, and objective comparisons use this fixed specification. The weights map faster recovery, greater probe readability relative to output retention, and stronger lightweight repair to higher accessibility.

B.3 RECOVERY-CURVE ESTIMATION

Anchor pools. MMLU: auxiliary-train domain split, 24k pool; IFEval: 8k self-built instruction-following examples spanning the 25 constraint types, generated and human-spot-checked; TruthfulQA: 6k paraphrase-generated true/false-adjudicated statements; AlpacaEval: 10k chat exchanges from open mixtures; safety: 5k refusal/compliance pairs from public red-teaming sets (evaluation items excluded); MGSM: 6k translated GSM-style items disjoint from test languages' items. All pools pass 8-gram de-duplication against every evaluation set (max collision rate 0.04%).

Paired normalization. For each capability, let $B_c(n)$ denote performance after fine-tuning the clean base θ_0 on n examples from its anchor pool using the Axis-1 LoRA protocol. We fit the reference curve by nonlinear least squares:

$$B_c(n) \approx A_c(\theta_0) + g_{\infty,c}\big(1 - e^{-n/\kappa_c^{\text{ref}}}\big), \qquad N_{50,c}^{\text{ref}} = \kappa_c^{\text{ref}} \ln 2,$$

where $g_{\infty,c} > 0$ is the fitted asymptotic improvement, so the reference half-time targets $A_c(\theta_0) + g_{\infty,c}/2$. The ratio in Eq. 2 compares recovery cost with this adaptation timescale; z-standardization then places it on the scale used by fusion.

Curve fitting. Nonlinear least squares on $\rho_c(n) = \rho_\infty(1 - e^{-n/\kappa})$ over $n \in \{0, 8, 32, 128, 512, 2048\}$, 3 seeds; $N_{50} = \kappa \ln \frac{\rho_\infty}{\rho_\infty - 0.5}$ when $\rho_\infty > 0.5$, else right-censored at

Table 5: Instance registry. Sub-pools of the same parent are disjoint unless marked $\subseteq$; the two verification sub-pools of the real-checkpoint set are drawn from disjoint FTS strata and together cover 39 of its 58 instances.

Pool	n	**Sampling frame**	**Parent / use**
Reference-labeled instances with externally established binary labels (total 186)			
Development suite (v1.0)	124	—	104 fitting + 20 construction-independence checks
$\subseteq$ mechanism-matched	104	5 recipes $\times$ 5 capabilities $\times$ 2 scales $\times$ 2 strengths, plus 4 replicates on the highest-seed-variance cells	**fusion weights fitted here**; AUROC 0.943 reported here
$\subseteq$ mechanism-agnostic	20	4 recipes, label from patching + replay ceiling only	construction-independence check; 87.5% agreement
Transfer set	62	—	fixed v1.0 applied
$\subseteq$ non-invertible	18	3 recipes with no approximate inverse in $\mathcal{R}^{\text{v1.0}}_{\text{light}}$	16/18 correct; AUROC 0.912 vs. 52 calibration overwritten references
$\subseteq$ cross-family	20	5 recipes on Llama-3.1-8B-Inst., 4 capabilities	AUROC 0.917
$\subseteq$ instruction-objective	24	5 recipes re-targeted at instruction/chat	AUROC 0.905; objective-specific recalibration
Continuous-label instances used for score-continuity analysis			
Mixed injections	20	S- and O-type recipes superposed at ratios $\{0.25, 0.5, 0.75\}$, labelled by realized ratio	FTS-continuity check only ($\rho = 0.86$); no binary label, hence outside the 186
Real post-training checkpoints and external validation			
Real-checkpoint pool	58	Qwen3-8B matched-gain, trajectory, and cascade checkpoints under both paradigms	39 externally verified: $36/39 = 92.3\%$, AUROC 0.936
FTS > 0.9	14	all instances in the two extreme strata	reverse validation, $22/24 = 91.7\%$
FTS < 0.1	10		
FTS $\in [0.3, 0.7]$ (mixed band)	19	of which 15 sampled, 5 per third of the band	mid-band dual verification; the 4 unsampled are unverified
FTS $\in [0.1, 0.3) \cup (0.7, 0.9]$	15	remainder; diagnosed but not externally verified	with the 4 above, the remaining diagnosed instances
Behavioral-signal strata	58	the same pool, partitioned at 0.3 and 0.7: 24/19/15	$\rho = 0.81$

4,096 and handled by rank statistics. Median fit $R^2 = 0.96$ (IQR $[0.93, 0.98]$). Table 7 reports the resulting shape separation, including the $11.5\times$ pair-normalized ratio of §3.2.

B.4 Probe Construction and Transfer

Probe tasks. MMLU: given the question and one option, classify correct/incorrect (balanced 50/50, 8k pairs); TruthfulQA: true/false statement discrimination (4k); IFEval: (instruction, response) constraint satisfied/violated (6k, violations synthesized by minimal edits); chat: preferred/dispreferred response discrimination from public preference data (6k); safety: harmful/benign request discrimination (4k, XSTest-style contrast pairs to control over-refusal shortcuts); MGSM: correct/incorrect solution discrimination (4k).

Probe fitting. Logistic regression uses the last-token residual stream. Nested five-fold question-level cross-validation selects the residual-stream layer and ℓ_2 penalty within each outer training

Table 6: The lightweight operator family $\mathcal{R}_{\text{light}}^{\text{v1.0}}$ and its validation search grids.

Operator	**Parameterization**	**Tuning within** B_{light}
Layer scaling (LiNeS-style)	$\theta_1^{(\lambda)} = \theta_0 + \Lambda \odot \delta$, $\Lambda_\ell = \lambda_{\min} + (\lambda_{\max} - \lambda_{\min}) \frac{\ell}{L}$ or free per-layer	Grid over $(\lambda_{\min}, \lambda_{\max}) \in \{0, 0.25, 0.5, 0.75, 1\}^2$; free per-layer via coordinate descent, 2 sweeps
Global interpolation	$\theta_\alpha = \theta_0 + \alpha\delta$, $\alpha \in [0, 1]$	Golden-section search, 8 evaluations
Format/prompt	Restore the base model's native chat template and system-prompt defaults; $k \in \{0, 2, 4\}$ few-shot format exemplars	Exhaustive over the 9 discrete combinations
Decoding	Temperature $\in \{0.0, 0.7, 1.0\}$; capability-cued prefix (fixed library of 6 prefixes)	Exhaustive over 18 combinations

Table 7: Recovery-curve statistics underlying Figure 3. N_{50} is measured in anchor examples; $\tilde{N}_{50}$ uses the paired reference curve (Eq. 2). Brackets report IQRs.

	median N_{50}	**median** $\tilde{N}_{50}$	**median** ρ_∞	**median** AUC-RC
Suppressed reference	18 [9, 41]	0.11 [0.06, 0.19]	0.97	0.88
Overwritten reference	312 [178, 704]	1.27 [0.81, 2.05]	0.83	0.41
Median ratio (overwr. / suppr.)	17.3×	**11.5**×	—	—
Mann–Whitney U		$p = 3.1 \times 10^{-9}$		

split, separately for θ_0 and θ_1. The selected probe is evaluated on the outer fold, and $P_c(\theta)$ averages the five held-out accuracies.

B.4.1 Probe Transfer Across Model Scales and Families

Within the Qwen3 family, the transfer branch applies the frozen 8B probe heads through a per-layer ridge correspondence map from the target residual stream into the 8B residual stream, fitted on 2k shared prompts (the family shares tokenizer and vocabulary); for architectures outside the family, the hybrid branch transfers the layer pre-ranking by relative depth and refits the probe head locally. With these branches, probes trained on Qwen3-8B reproduce locally retrained G_c rankings at Spearman $\rho = 0.76$, 0.81 and 0.79 on Qwen3-1.7B, 4B and 14B, with Axis-2 AUROC differences of 2.8%, 1.9% and 2.2%: pooled Qwen $\rho = 0.78$ at a 2.4% difference, so all three take the transfer branch for within-family §4.5. Llama-3.1-8B-Instruct falls in the hybrid band ($\rho = 0.71$, 3.4%), and §4.5 uses the hybrid strategy there—pre-ranking probes transferred, the final one retrained locally—shifting typing consistency by 1.8% vs. full local retraining. Worst-case full local retraining costs < 0.1 GPU·h per model; the all-models retraining ceiling for our study is 0.7 GPU·h total.

B.5 Two-Stage Repair Algorithm

Router resolve step. Global operators (layer scaling, interpolation) are selected jointly from the cached Axis-3 candidate grid using the union of the validation slices of the capabilities that receive a light operator (weighted by Δ_c); local operators (prompt/format, decoding) stack independently per capability. This selection requires no new operator search. Conflicts (a global coefficient helping c_1 but hurting c_2 beyond 0.5 points) occurred in 2 of 30 routed runs and were resolved by per-capability prompt-level overrides.

B.6 End-to-End Diagnostic and Repair Protocol

This subsection summarizes one complete TRIAGE run, separating the one-time calibration from per-checkpoint deployment.

Phase 0: offline calibration (once per score version). (i) Build the anchor pools and probe tasks per capability (Appendices B.3, B.4); (ii) fit the seven reference recovery curves on the clean base θ_0

Algorithm 1 TRIAGE-Router (§3.4)

Require: θ_0, θ_1, capability set $\mathcal{C}$, total replay budget B, target recovery $\bar{\rho}$, thresholds τ_{hi}=0.7, τ_{lo}=0.3
Ensure: repaired model θ^*
$\mathcal{D} \leftarrow \{c \in \mathcal{C} : \Delta_c \geq \tau_{\min}\}$
for $c \in \mathcal{D}$ **do**
  compute FTS_c and cache candidate scores/configurations {Axis-3 search is paid once}
end for
$\mathcal{S} \leftarrow \{c \in \mathcal{D} : \text{FTS}_c \geq \tau_{\text{hi}}\}$; $\mathcal{O} \leftarrow \{c \in \mathcal{D} : \text{FTS}_c \leq \tau_{\text{lo}}\}$; $\mathcal{M} \leftarrow \mathcal{D} \setminus (\mathcal{S} \cup \mathcal{O})$
move A1/A3-flagged capabilities from $\mathcal{S}$ or $\mathcal{O}$ to $\mathcal{M}$ {routing-tier override}
Stage 1 (suppressed and mixed → light repair, near-zero cost, gradient-free):
$\theta' \leftarrow \text{resolve}(\{r_c : c \in \mathcal{S} \cup \mathcal{M}\})$ {global candidates selected from cache; local operators reused}
re-evaluate on held-out data; retain if recovery $\geq \eta$ for $c \in \mathcal{S}$ or > 0 for $c \in \mathcal{M}$, and retention holds
Stage 2 (overwritten + mixed → targeted replay, continuous allocation):
$n_c \propto (1 - \text{FTS}_c) \cdot \Delta_c$ for $c \in \mathcal{O} \cup \mathcal{M}$, $\sum_c n_c \leq B$ {no hard threshold dependence}
$\theta^* \leftarrow \text{SFT}\big(\theta', \bigcup_c D_c[n_c] \cup \text{task-data mix}, +\text{KL}(\cdot\|\theta') \text{ on task prompts}\big)$
verify on held-out dev; if mean recovery is below $\bar{\rho}$ and budget remains, run one additional allocation round
return θ^*

under the Axis-1 protocol (25.4 GPU·h for all fourteen Axis-1 calibration curves; reference curves shared by all checkpoints of that base); (iii) construct the controlled reference states and label them with the patching–replay rule of §3.3; (iv) fit the logistic weights and standardization statistics on the 104 mechanism-matched references and freeze them as v1.0 (Appendix B.2).

Phase 1: per-checkpoint diagnosis (5.9 GPU·h with shared reference curves; 9.8 for the first checkpoint of a new base). (i) Gate capabilities at $\Delta_c \geq 2$ on the three-seed mean; (ii) Axis 1: run the deployment grid $\{0, 6, 24, 96, 384\}$ as independent prefix-nested LoRA fine-tunes of θ_1 over three seeds, fit $\rho_c(n)$, and normalize by the shared reference curve, with right-censored curves entering at the 4,096 cap; (iii) Axis 2: evaluate the transferred or locally refit probe on θ_0 and θ_1 and form G_c; (iv) Axis 3: sweep $\mathcal{R}^{\text{v1.0}}_{\text{light}}$ within $B_{\text{light}} = 64$ validation examples under the retention constraint, caching validation scores and the maximizing configuration; (v) fuse the three standardized axes with the frozen v1.0 statistics into FTS_c and assign tiers, with reliability flags promoting flagged capabilities to the mixed band.

Phase 2: repair and verification. Stage 1 selects global operators jointly from the Axis-3 cache and stacks per-capability prompt and decoding configurations, subject to the held-out acceptance gate; Stage 2 allocates replay in proportion to $(1 - \text{FTS}_c)\Delta_c$ from the Stage-1 model under the KL safeguard, verifies recovery and retention on held-out data, and optionally runs one additional allocation round. The output is a single repaired checkpoint θ^* together with the per-capability inference-time configuration sheet; itemized costs follow Appendix E.3.

C Reference Construction and Calibration Data

C.1 Mechanism-Controlled Reference Recipes

All injections start from Qwen3-8B or Qwen3-1.7B (×2 scales) and target one of five capabilities (knowledge/MMLU, factuality/TruthfulQA, instruction following/IFEval, chat/AlpacaEval, safety/AdvBench+XSTest), at two strengths (mild: target $\Delta_c \in [3, 6]$ points; severe: $\Delta_c \in [8, 15]$). Table 8 summarizes the five recipes and their literature anchors.

- **S-i (format distortion).** SFT on 4k examples whose responses are semantically correct but wrapped in a distorted output schema (JSON-of-wrong-keys, forced prefix boilerplate); LR 5×10^{-6}, 1 epoch. Verified to leave MMLU log-prob margins on gold options within 0.5% of θ_0.

Table 8: Mechanism-matched injections: each recipe is a controlled amplification of a forgetting mechanism localized in the literature.

Injection	Mechanism matched	Literature anchor
S-i: format/template-distortion training only	Post-training changes the output-format mapping, not knowledge	Yun et al. (2025); Lin et al. (2024)
S-ii: very-low-LR shallow task alignment	Task-alignment drift masquerading as knowledge loss	Zheng et al. (2025) (original recipe reproduced)
S-iii: alignment shift trained on top-N layers only	Alignment updates concentrate in shallow layers / early tokens	Qi et al. (2025)
O-iv: counterfactual-corpus high-LR multi-epoch overwrite	Long SFT overwriting factual representations	knowledge-editing literature (Meng et al., 2022)
O-v: gradient-ascent unlearning on anchor data	Targeted representation destruction	unlearning literature (Jang et al., 2023)

- **S-ii (shallow task alignment).** Reproduction of the spurious-forgetting recipe: LR 2×10^{-6}, 200 steps on task data, updates dominated by the first 3 transformer blocks (verified by per-layer update norms).
- **S-iii (top-layer alignment shift).** Full-LR training with all but the top $N \in \{4, 8\}$ layers frozen, on alignment-style data with shifted refusal/format policy. At severe strength the reference measurement places most of these cells on the overwritten side (§4.1).
- **O-iv (counterfactual overwrite).** SFT on 12k counterfactual statements contradicting anchor facts, LR 2×10^{-5}, 3 epochs—the standard knowledge-overwrite regime.
- **O-v (gradient-ascent unlearning).** Ascent on anchor-capability data with KL retention on unrelated text, 150 steps, following standard unlearning practice.

C.2 Mixed Injections

20 instances superpose S-type and O-type recipes at mixing ratios $\{0.25, 0.5, 0.75\}$ (by interleaved step counts), labeled with the realized ratio; used only for the FTS-continuity check (Spearman $\rho = 0.86$).

C.3 Construction-Independent Reference Recipes

Four additional recipes are chosen without using any diagnostic-axis value: (i) random-layer-subset perturbation (uniformly sampled 25% of layers, LR 3×10^{-5}, 300 steps on task data); (ii) task-vector random low-rank projection removal (remove a random rank-k subspace, $k \in \{8, 64\}$, of δ); (iii) interference training on an unrelated task (code SFT) to induce collateral damage; (iv) targeted embedding noise (Gaussian, σ scaled to 10% of row norms, on 5% of vocabulary rows tied to the capability domain). Reference labels for these instances use causal patching and the replay ceiling rather than the intended effect of the recipe.

C.4 Cross-Family References

To construct a reference-labeled transfer set outside Qwen3, we re-ran the five mechanism-matched recipes of Table 8 on Llama-3.1-8B-Instruct, at one scale over four capabilities (knowledge, factuality, instruction following, safety), with mild and severe strengths balanced across recipes—one strength per recipe–capability pair—yielding 20 held-out instances. Recipe hyperparameters were adjusted only where the architecture requires it (chat-template tokens for S-i, layer indices for S-iii); the label-defining protocol—causal patching plus replay ceiling—is unchanged. Frozen FTS v1.0 types these instances at AUROC 0.917 $[0.851, 0.971]$ with $18/20$ correct, and the two misses are a mild S-iii instance on safety (FTS 0.63, inside the mixed band) and an O-iv instance on factuality whose replay ceiling (0.71) sits above the overwritten-typical range. The format/prompt operator uses Llama's native template under the model-native v1.0 rule (Appendix B.1); an intentionally non-native Qwen-template stress test lowers AUROC to 0.884.

Table 9: Diagnostic performance. AUROC is threshold-free; F1 and accuracy use threshold 0.5, while routing retains the $[0.3, 0.7]$ mixed band. The top block evaluates the reference set (LOIO = leave one injection type out); the bottom block evaluates transfer. Bootstrap 95% CIs; CKA baseline per Xu et al. (2026).

Criterion	n	**AUROC**	**F1**	**Acc.**
Axis 1 only: $z(-\log \tilde{N}_{50})$	104	0.891 $[0.848, 0.927]$	0.858	86.5%
Axis 2 only: $z(G_c)$	104	0.862 $[0.815, 0.904]$	0.831	83.7%
Axis 3 only: $z(\mathrm{LRR})$	104	0.908 $[0.869, 0.941]$	0.877	88.5%
CKA representation-similarity baseline	104	0.716 $[0.652, 0.778]$	0.694	70.2%
Sample-wise forgetting ($1 \to 0$ rate)	104	0.760 $[0.678, 0.850]$	0.705	71.2%
Behavioral error-type signal (no axis used)	104	0.861 $[0.795, 0.918]$	0.830	84.6%
Fused FTS (v1.0)	104	$\mathbf{0.943}$ $[0.914, 0.968]$	$\mathbf{0.921}$	$\mathbf{92.3}\%$
FTS, LOIO generalization (mean $\pm$ s.d.)	104	0.921 ± 0.018	0.897 ± 0.021	90.1%
Non-invertible mechanisms[†]	70	0.912 $[0.842, 0.968]$	—	—
Llama-3.1-8B-Inst. transfer	20	0.917 $[0.851, 0.971]$	0.895	90.0%
Instruction-objective transfer	24	0.905 $[0.838, 0.960]$	0.873	87.5%

C.5 Suppressions Without Available Operator Inverses

Three suppression recipes without structural inverses in $\mathcal{R}^{\mathrm{v1.0}}_{\mathrm{light}}$ produce 18 held-out instances (3 recipes $\times$ 3 capabilities $\times$ 2 strengths). *N-i, unembedding-adjacent rotation*: train a norm-preserving rotation of the final-layer readout subspace, leaving mid-stack features untouched; the Axis-3 maximizer is layer scaling $(5/6)$ or decoding $(1/6)$. *N-ii, answer-withholding persona*: SFT on a persona that hedges or defers on the target capability while retaining competence; repaired by decoding/prefix $(4/6)$ or interpolation $(2/6)$. *N-iii, length/budget collapse*: SFT toward truncated responses so the capability is cut off rather than lost; repaired by decoding $(3/6)$, format $(2/6)$ or interpolation $(1/6)$.

All 18 instances carry suppressed reference labels established by causal patching plus replay ceiling, independently of recipe intent; the calibration overwritten instances provide the negative pool for metrics that require one. Frozen FTS types $16/18$ correctly (AUROC 0.912, Table 9). The two boundary instances are severe N-i rotations on knowledge, where large readout-space rotation reduces probe readability together with output accuracy and moves Axis 2 toward the overwritten pole. In $13/18$ instances the repairing operator is not the mechanism's structural counterpart: rotations are absorbed by layer scaling, personas by prefix steering, and truncation by decoding changes. These cases extend the diagnosis to recovery paths beyond direct reversal of the injected perturbation.

D Diagnostic Validation and Sensitivity

D.1 Fusion Analysis

[†]AUROC uses 18 non-invertible suppression cases and 52 calibration overwritten references; FTS correctly types $16/18$ transfer cases.

Diagnostic comparisons. Table 9 reports calibration, LOIO, and transfer performance. Fusion improves AUROC by 0.052, 0.081, and 0.035 over Axes 1–3, respectively (DeLong $p = 0.008$ for the smallest gain). The sample-wise forgetting baseline reaches 0.760 $[0.678, 0.850]$ on the reference set and 0.719 ± 0.053 under LOIO, with fold values 0.650–0.790. Transfer-set construction appears in Appendices C.5, C.4, and I.1.

Input complementarity and fitted weights. Pairwise correlations of the standardized inputs are 0.41 (Axes 1–2), 0.63 (1–3), and 0.47 (2–3). Variance-inflation factors are 1.9, 1.4, and 2.1, below the ridge-fallback threshold of 5. Across 500 bootstrap resamples, the 95% intervals are $w_1 \in [1.42, 2.31]$, $w_2 \in [0.58, 1.39]$, $w_3 \in [0.94, 1.85]$, and $b \in [-0.55, 0.12]$; the ordering $w_1 > w_3 > w_2$ holds in 91% of resamples.

Axis attenuation and low-rank damage. Shrinking one axis's standardized value toward zero by $\gamma = 0.75/0.5/0.25$ gives AUROC $0.930/0.912/0.887$, from 0.943, for the most affected axis (Axis 1). On 18 low-rank, above-median-norm damage instances, FTS reaches AUROC 0.91, compared with 0.58 for CKA.

D.2 External Verification on Real Checkpoints

We externally evaluate 24 real-checkpoint instances from the two extreme FTS strata; 22 agree with the corresponding reference verdict. For each instance, we compute (1) width-4 causal-patching recovery p_c; (2) 4,096-example, two-epoch replay-ceiling recovery u_c; and (3) DTW distance between its normalized $\rho_c(\log n)$ curve and the two calibration-pole templates. A predicted suppression-like instance agrees when $p_c \geq 0.70$, $u_c \geq 0.50$, and its curve is closer to the suppression template; the overwrite-like verdict uses the complementary inequalities. Discordant measurements are retained as boundary cases, with DTW used as corroborating evidence. Table 10 reports the outcomes.

Table 10: Reverse-validation outcomes (the 24 extreme-FTS instances of the 58-instance real-checkpoint pool; sampling frame in Table 5).

Group	n	**Pass**	**Boundary cases (§D.9)**
FTS > 0.9 (predicted suppressed)	14	13	1 (patching window found only at width 8)
FTS < 0.1 (predicted overwritten)	10	9	1 (replay ceiling reached 0.62 recovery)
Total	24	22 (91.7%)	2

D.3 Probe Readability and Causal Recovery

Probe evaluation. Under five-fold question-level cross-validation, selectivity over shuffled-label controls exceeds 18 points for every capability. Adversarial paraphrases reduce accuracy by less than 2.5 points, and probe-task difficulty correlates with original-task difficulty at $r = 0.42$.

Causal patching check (calibration only). Patching θ_0 activations into θ_1 at the probe-optimal layer window: output recovery correlates with PRR at $r = 0.74$ across the calibration set; among instances with $\mathrm{PRR} > 0.9$, 88% show $\geq 60\%$ output recovery under patching.

Patching across real post-training checkpoints. We apply the same patching protocol to all 58 real-checkpoint instances (§4.2) and stratify by damage effective rank. In the high-rank stratum ($n = 51$), PRR correlates with patch recovery at $r = 0.71$ $[0.55, 0.82]$; the relation remains $r = 0.66$ among its 15 mixed-band instances. The low-rank stratum ($n = 7$) is less reliable ($r = 0.52$ and median patch recovery 41%, versus 74% at comparable PRR in the high-rank stratum). These cases all enter the mixed-confidence tier through either the A1 flag or their FTS value. Incremental cost is ≈ 0.2 GPU·h.

From readability to behavior. PRR measures linear decodability, and patching connects this measurement to restored outputs. The correlation is strongest in the high-rank stratum; high-PRR/low-patching cases enter the mixed-confidence tier through the A1 rule. The fitted fusion weights satisfy $w_2 < w_3 < w_1$.

D.4 Behavioral Corroboration

The model's errors provide a behavioral view of accessibility. We measure whether newly failed items change primarily in surface form or in content, and compare this output-based signal with FTS.

Protocol. For each diagnosed instance we collect the items that θ_0 answered correctly and θ_1 does not, and classify each failure into *surface-form* (constraint/format violation, refusal or over-refusal, truncation, language slip, answer present but unparseable) or *content* (factually wrong, wrong reasoning step, wrong option selected with a coherent but incorrect rationale). Classification uses a fixed rubric applied by two independent LLM judges with a human adjudication pass on disagreements (judge agreement $\kappa = 0.78$; 180 adjudicated items). The signal is the surface-form share $S_c \in [0, 1]$ of newly-failed items. Cost: no GPU training, ≈ 0.1 GPU·h of inference per instance.

Prediction. Suppression-like loss should concentrate errors in how an answer is delivered; overwrite-like loss should concentrate them in what the answer says.

Table 11: Surface-form share S_c of newly failed items across the 58 real-checkpoint instances and 104 calibration references. Real-checkpoint rows use FTS-assigned types; calibration rows use reference labels.

Group	n	Surface-form share S_c	Content share
Real checkpoints, FTS-suppressed (FTS > 0.7)	24	71.4% $[64.8, 77.6]$	28.6%
Real checkpoints, mixed band	19	46.2% $[38.1, 54.7]$	53.8%
Real checkpoints, FTS-overwritten (FTS < 0.3)	15	18.9% $[13.2, 26.1]$	81.1%
Calibration, suppressed reference labels	52	74.8%	25.2%
Calibration, overwritten reference labels	52	16.3%	83.7%

Results. Surface-form errors increase monotonically across overwrite-dominant, mixed, and suppression-dominant losses (Table 11). Across 58 real instances, S_c correlates with FTS at Spearman $\rho = 0.81$ $[0.70, 0.88]$ and reaches standalone AUROC 0.861 on the calibration set. The output pattern complements the recovery and representation measurements: accessible losses predominantly alter expression, while less accessible losses predominantly alter content.

D.5 External Verification in the Mixed Band

We complement the extreme-score evaluation with 15 real-checkpoint instances sampled from FTS $\in [0.3, 0.7]$: five per third of the band, spanning both paradigms and four capabilities. Each instance receives the same patching and replay-ceiling measurements.

Graded verification score. We average two normalized measurements, $V = (v_p + v_u)/2 \in [0, 1]$. Here v_p is patch recovery divided by 0.7 and clipped to $[0, 1]$; v_u is replay recovery relative to the suppressed-typical ceiling, also clipped to $[0, 1]$. Both components are fixed before sampling; higher V indicates recovery closer to the suppression references.

Findings. V rises across the three FTS sub-bands: means are 0.41, 0.46, and 0.58 on $[0.30, 0.43)$, $[0.43, 0.57)$, and $[0.57, 0.70]$ (five instances each), giving Pearson $r = 0.61$ $[0.14, 0.85]$ over the 15 instances. Eleven are intermediate under both external measurements, three agree with the direction of the FTS lean, and one disagrees. The disagreement is an A1-flagged MGSM loss under SFT, consistent with the low-rank limitation identified in Appendix B.4. These results support treating the band as graded uncertainty. Because Stage 2 allocates replay continuously in $(1 - \mathrm{FTS}_c)$, such uncertainty produces a proportional budget change rather than an inverted binary decision (Appendix H.2).

D.6 Label Sensitivity of the Reference Measurement

We vary the patching window, patch-recovery threshold, and replay-ceiling budget in the reference rule of §3.3, keeping the replay-recovery split at $u_c = 0.50$. Table 12 scores fixed FTS v1.0 on the concordant subset of the 104 references under each protocol. "Flips" counts assignments that change or become discordant. This analysis evaluates reference-label agreement; downstream indices continue to use the fixed FTS values.

The largest change is 1.9 AUROC points and 9 of 104 reference assignments. FTS remains above 0.9 AUROC throughout. Widening the patching window produces the largest single change, matching the distributed-suppression case in Table 10.

Held-out operator evaluation. Re-evaluating each selected operator on a disjoint 64-example slice changes mean recovery from 0.71 to 0.66, Axis-3 AUROC from 0.908 to 0.895, and fused FTS AUROC from 0.943 to 0.936. Stage 1 uses a held-out evaluation of the selected configuration before repair acceptance.

Table 12: Label sensitivity under alternative reference-measurement thresholds. Frozen FTS v1.0 is scored on the concordant subset induced by each protocol; "Flips" counts assignments that change or become discordant relative to the default.

Threshold varied	Setting	Flips	FTS AUROC	Axis-3 AUROC
— (default protocol)	width 4, 70%, 4k	—	0.943	0.908
Patching window	width 8	6	0.929	0.914
Patching recovery	60%	4	0.936	0.899
Patching recovery	80%	7	0.931	0.912
Replay ceiling budget	8k examples, 2 ep	3	0.939	0.905
Both loosened	width 8, 60%	9	0.924	0.903

D.7 Recovery-Curve Sensitivity

Reference-curve stability across anchor pools. For each capability, we recompute the reference curve on θ_0 using a 50% subsample, the disjoint second half, and an alternative-source pool. The resulting $\tilde{N}_{50}$ rankings correlate at Spearman $\rho = 0.86$–0.93 per capability (median 0.89), while the suppression-like/overwrite-like median ratio remains $10.3\times$–$12.6\times$. MGSM has the lowest correlation (0.86) and is flagged by the A3 reliability rule. The fallback for correlations below 0.85 uses a cross-dataset estimate of intrinsic learning difficulty.

Table 13: Kendall-τ of $\tilde{N}_{50}$ rankings against the reference protocol (LR $1\times$, rank 16), and the suppressed/overwritten median ratio, per grid cell (32-instance subset).

	LoRA $r = 8$		LoRA $r = 16$		LoRA $r = 32$	
LR	τ	ratio	τ	ratio	τ	ratio
$0.5\times$	0.86	$9.4\times$	0.89	$10.8\times$	0.87	$8.7\times$
$1\times$	0.88	$11.1\times$	— (ref)	$11.9\times$	0.91	$13.2\times$
$2\times$	0.86	$9.1\times$	0.90	$12.4\times$	0.88	$10.6\times$

Across the optimization grid, the median separation ratio exceeds $8\times$ and ranking agreement has median $\tau = 0.88$ (minimum 0.86). Comparing one versus two epochs gives $\tau = 0.90$. All settings satisfy the A2 stability threshold in Table 14.

Budget-grid sensitivity. On 16 instances balanced by reference type and spanning all five recipes, a shifted grid $\{0, 16, 64, 256, 1024\}$ and the deployment grid $\{0, 6, 24, 96, 384\}$ yield Kendall-$\tau = 0.91$ and 0.86 against the full calibration grid. Their median type-separation ratios are $10.6\times$–$12.3\times$.

Right-censored curves. Eleven of 104 references (10.6%), all overwrite-labeled, have fitted $\rho_\infty < 0.5$. Their N_{50} is right-censored at 4,096 and handled by rank statistics. Excluding these instances changes the median separation ratio from $11.5\times$ to $10.9\times$ and leaves typing AUROC unchanged to three decimals.

D.8 Operator-Family Scope and Sensitivity

Cross-family portability. The same operator semantics and search budgets are used for every model, while prompt restoration uses each model's native chat template and system defaults. On Llama-3.1-8B-Instruct, recovery on suppression-dominant losses remains close to Qwen3 for layer scaling ($86.2\% \rightarrow 84.1\%$), global interpolation ($83.5\% \rightarrow 82.7\%$), and decoding interventions ($79.8\% \rightarrow 76.5\%$). Native prompt restoration reaches 85.3%; reusing the Qwen template instead reaches 71.4%, showing that the transferable object is the restoration rule rather than a specific token sequence.

Removing layer scaling, global interpolation, format/prompt restoration, or decoding gives Kendall-$\tau = 0.93$, 0.95, 0.94, and 0.97 against the full family. Adding SVD low-rank projection gives $\tau = 0.96$ and changes the type of 3.2% of capabilities. The accessibility ranking is stable across these operator-family variants.

D.9 Boundary Cases and Conflicting Evidence

Reliability-aware weighting. Table 14 summarizes three conditions under which an axis may be less reliable and should receive lower confidence. A1 identifies large-norm, low-rank updates; A2 detects unstable recovery-curve rankings across optimization choices; and A3 detects unstable reference normalization. Verification agreement is 84% among A1-flagged instances and 93% among unflagged instances. In the reported experiments, A1 flags the relevant low-rank cases, A3 fires for MGSM at 1.7B, and A2 is not activated.

Table 14: Reliability rules: measurement conditions, triggers, routing responses, and observed frequencies. Appendix D.7 reports the A2 optimization grid.

Assumption	Content and main risk	Trigger and weighting rule	Observed
A1 (low-dim. readout)	Suppressed recovery corresponds to low-dimensional readout re-alignment; risk: large-norm damage concentrated at low rank can fake "low-dim. fast recovery"	Damage effective rank (task-vector spectrum, energy-90% rank) below the calibration 25th percentile *and* $\|\delta\|$ in the calibration top quartile $\Rightarrow$ instance flagged *A1-risky*: FTS retained, confidence demoted one tier, routed to the mixed band	$9/104$ calibration (8.6%) and $4/58$ real instances flagged; verification agreement 84% among flagged vs. 93% unflagged
A2 (shape–type map stable within protocol)	Under a fixed fine-tuning protocol, curve shape is determined mainly by representation integrity; risk: shape sensitive to LR/LoRA rank	Grid Kendall-τ of $\tilde{N}_{50}$ rankings $\in [0.70, 0.85) \Rightarrow w_1 \leftarrow w_1 \cdot \frac{\tau - 0.70}{0.85 - 0.70}$, released weight mass renormalized to $w_2{:}w_3$; $\tau < 0.70 \Rightarrow w_1 = 0$ and a two-axis refit, with before/after weights and AUROC reported	$\tau = 0.88$ (grid range 0.86–0.91): rule not triggered
A3 (difficulty confounds normalizable)	Intrinsic capability difficulty can be removed by the paired baseline of Eq. 2; remaining variance after normalization	Within-capability seed variance of $\tilde{N}_{50}$ exceeding $z = 1.5$ against the calibration distribution $\Rightarrow$ capability enters the mixed-confidence tier; composition results reported stratified	Triggered for MGSM at 1.7B only

Assumption-signature map. The A1 signature is low $\tilde{N}_{50}$ (suppression-like) paired with PRR collapsing alongside ORR, non-restorative causal patching, and a low-rank damage spectrum with a large update norm. The A2 signature is a $\tilde{N}_{50}$ ranking change across LR/LoRA-rank settings or a shape tier crossing the calibration grading boundary. The A3 signature is unusually large seed variance in N_{50}^{ref} or a post-normalization $\tilde{N}_{50}$ rank correlation with Axes 2–3 below that of the other capabilities.

The five calibration conflicts (5/104). Three are nearest neighbors of mixed injections (S-iii severe on safety at 1.7B; O-iv mild on chat at both scales): FTS lands in $[0.42, 0.58]$, and the two verification methods also disagree, identifying genuinely mixed cases for which the continuous score is the appropriate output. Two carry the A1 low-rank signature: Axis 1 leans suppressed while Axes 2–3 and causal patching lean overwritten, exactly the pattern captured by the A1 flag.

The two reverse-validation boundary cases ($2/24$ of Table 10; protocol in Appendix D.2). One predicted-suppressed instance (IFEval, GRPO checkpoint 2) requires a width-8 patching window rather than the criterion's width 4, indicating distributed suppression across adjacent layers. One predicted-overwritten instance (MMLU-Pro under severe SFT) reaches 0.62 replay-ceiling recovery, above the 0.5 overwritten-reference threshold and below suppressed-typical levels, placing it near the mixed band. Both occur at the band boundary and are localized by the criterion's confidence machinery.

Construction-independent boundary cases (2.5/20 = 12.5%). The half-scored ambiguous instance is random low-rank projection removal ($k = 64$) on knowledge: FTS $= 0.51$ lies in the mixed band, and the two verification methods diverge (patching restores 58%, replay ceiling 0.74), motivating the half score. The first full disagreement is interference training (code-SFT collateral damage on IFEval): FTS $= 0.38$ while verification leans suppressed (patching restores 74%); its low-rank spectrum triggers the A1 flag. The second is targeted embedding noise on factuality: FTS $= 0.64$ while verification leans overwritten (replay ceiling 0.41); vocabulary-local noise preserves most probe features and raises Axis 2. All three lie in $[0.3, 0.7]$, so disagreement is concentrated in the mixed band and A1-flagged region identified by the confidence machinery.

E TRAINING, DIAGNOSIS, AND COST HYPERPARAMETERS

E.1 POST-TRAINING CONFIGURATIONS

Table 15 specifies the main SFT and GRPO runs.

Table 15: Post-training configurations (main runs).

	SFT	**GRPO (RLVR)**
Data	NuminaMath + OpenR1-Math, 210k	DeepScaleR prompts, 40k
Epochs / steps	2 epochs (3 for trajectory)	$\sim$800 steps (1,200 for trajectory)
LR	1×10^{-5}, cosine, 3% warmup	actor 1×10^{-6}, constant
Batch	global 256, seq len 4,096	256 prompts $\times$ 8 rollouts
KL to reference	— ($\beta = 0.01$ in factorial arms)	$\beta = 0.01$; sweep $\{10^{-3}, 10^{-2}, 10^{-1}\}$, β=0 control
Reward	—	binary verifiable (boxed-answer match)
Framework	LLaMA-Factory	verl (Sheng et al., 2025) (OpenRLHF-style stack; cf. Hu et al., 2025)
Hardware	8×H800 (16 for 14B)	8×H800 (16 for 14B)

E.2 DIAGNOSTIC AND REPLAY SETTINGS

Axis-1 recovery runs. LoRA $r = 16$, $\alpha = 32$, dropout 0; LR 1×10^{-5}; 1–2 epochs; batch 32; 3 seeds. The calibration grid is $n \in \{0, 8, 32, 128, 512, 2048\}$; the deployment grid is the compressed $\{0, 6, 24, 96, 384\}$ of Appendix D.7. Reference curves on θ_0 use the identical protocol; per-curve costs are derived below.

Axis-2 probes. Logistic regression on the last-token residual stream at each layer; nested 5-fold question-level CV, with layer and ℓ_2 regularization selected on each outer fold's training split separately for θ_0 and θ_1, then evaluated on the untouched outer fold; chance normalization $\tilde{P} = (P - P_{\text{chance}})/(1 - P_{\text{chance}})$.

Axis-3 / router. $B_{\text{light}} = 64$; $\eta = 0.8$, $\eta_{\text{low}} = 0.3$, $t_N = 0.5$; $\tau_{\text{hi}} = 0.7$, $\tau_{\text{lo}} = 0.3$; Stage 2 mixes 15% new-task data; KL-to-θ' weight 0.1; dev gate 98%; rollback halves LR, with at most one additional allocation round.

E.3 COMPUTE ACCOUNTING

Measured throughput. Diagnosis and replay are priced using their measured per-sample throughput. At the common $\approx 90\%$ target, uniform replay consumes 96k examples in 73.8 GPU·h (Table 26): 0.77 GPU·h per 1k examples for full-parameter updates with a KL reference forward pass. Axis-1 curves are LoRA $r = 16$ fits with no reference model and no optimizer state for the frozen weights, measured at $4.3\times$ that throughput, i.e. 0.18 GPU·h per 1k examples. Evaluation is capability-scoped (the diagnosed benchmark only, not the full suite) at 0.019 GPU·h per point.

Calibration grid. One curve visits $2{,}728$ examples per seed, $8{,}184$ over three seeds: 1.47 GPU·h of fitting plus 18 evaluation points at 0.019, so 1.8 GPU·h per curve. Establishing FTS v1.0 required this grid for both the seven target and the seven reference curves, 25.4 GPU·h in total. This is the one-time calibration cost for the target and reference curves.

Deployment grid. The lower-cost setting uses the same compressed budget grid for the target and reference curves. One curve costs 510 examples per seed, or $1{,}530$ over three seeds: 0.275 GPU·h of fitting plus 15 evaluation points, for 0.56 GPU·h per curve and 3.9 for seven. Its $\tilde{N}_{50}$ rankings agree with the full calibration grid at Kendall-$\tau = 0.86$. Re-running the closed loop with full-grid Axis-1 values raises recovery from 91.7% to 92.3% at an additional 8.8 GPU·h, without changing the method ordering.

Other axes. Axis 2: linear probes over 36 layers $\times$ 7 capabilities on cached activations, 0.5 GPU·h, dominated by the activation dump rather than the probe fits. Axis 3: every configuration in the $\mathcal{R}^{\text{v1.0}}_{\text{light}}$ grid tuned within B_{light} on the validation slice, each followed by a retention check, 1.5 GPU·h. Per-

checkpoint total $3.9 + 0.5 + 1.5 = 5.9$ GPU·h, i.e. ≈3.7% of the 161 GPU·h SFT run it diagnoses (210k examples at the replay throughput above).

First-checkpoint cost. Reference curves are shared across checkpoints from the same base model. Computing all seven references for a new base raises diagnosis cost from 5.9 to 9.8 GPU·h. The corresponding savings over uniform replay are $2.0\times$ for RLVR-degraded checkpoints, $1.6\times$ for SFT-degraded checkpoints, and $1.7\times$ pooled.

F Detailed Forgetting Anatomy Across Capabilities and Models

F.1 Per-Capability Full Results

Table 16 connects measured KL to the individual diagnostic axes. The subsequent trajectory and model-scale results expand the composition comparison of §4.2.

Table 16: Relation between measured KL and the diagnostic axes (26 points: 14 matched-gain cells and 12 KL-sweep runs). The first two rows report individual-axis measurements; FTS and training steps provide complementary comparisons.

Signal	Predicted direction vs. KL	Observed ρ	p	Verdict
$\tilde{N}_{50}$ (Axis 1, pre-fusion)	positive	$+0.63$	0.0021	aligned
G_c (Axis 2, pre-fusion)	negative	-0.57	0.0058	aligned
FTS (auxiliary, fused)	negative	-0.61	0.0034	consistent
Steps at fixed KL (control)	none if KL-driven	$+0.14$	0.41	separable

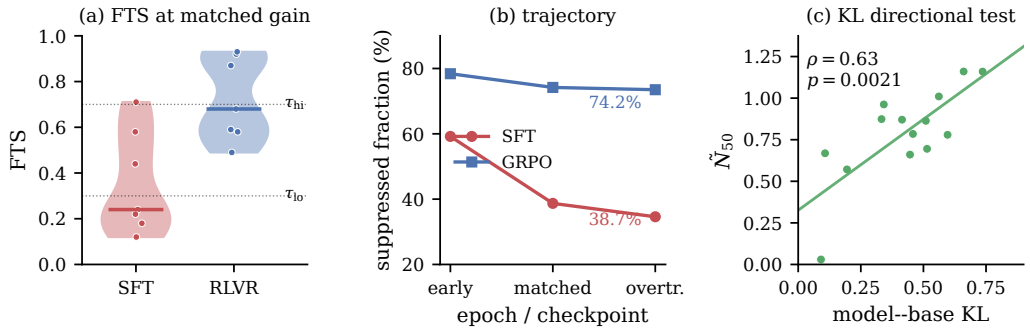

Figure 4: SFT and RLVR forgetting anatomy. Recovery states evolve along training trajectories, and the individual diagnostic axes track measured KL to the base model.

Table 17: Trajectory-resolved suppression indices (Qwen3-8B; FTS-weighted % of loss mass, mixed-band mass split proportionally to FTS as in Appendix J). Capability-clustered bootstrap 95% CIs.

Checkpoint	SFT	GRPO	SFT→GRPO cascade
Early (SFT ep. 1 / GRPO step 250)	59.2 [47.4, 71.0]	78.4 [67.5, 87.6]	—
Mid (SFT ep. 1.5 / GRPO step 500)	45.3 [34.5, 56.5]	74.9 [64.1, 84.5]	—
Matched-gain stop (ep. 2 / step 800)	38.7 [28.7, 49.1]	74.2 [63.4, 83.9]	44.1 [32.8, 55.7]
Overtrained (ep. 3 / step 1200)	34.6 [24.6, 45.4]	73.5 [62.5, 83.5]	—

Matched gain and compute exposure (SFT epoch-1 control). The matched-gain stopping points of §4.2 equate the paradigms on task outcome: the SFT stop corresponds to ≈ 1,640 gradient updates (2 epochs at global batch 256) versus ≈ 800 policy updates for GRPO, each preceded by 8-rollout sampling. Two controls characterize this difference. *First*, SFT stopped at epoch 1 (GSM8K +8.9 / MATH500 +7.1, ≈ 75% of final gain) yields a suppression index of 59.2% (Table 17, Early row),

between GRPO's 74.2% and the 38.7% reached after epoch 2. Most additional SFT overwriting therefore accrues during the second epoch, after task loss has largely saturated. *Second*, a denser GRPO trajectory (checkpoints every 100 steps, 8 points) keeps the suppression index within 73–79% (s.d. 1.9 points, no monotone trend), confirming the flat trajectory in Figure 4b. Matched task gain supplies the practitioner-relevant anchor, while the epoch-1 comparison quantifies the effect of additional SFT exposure.

Severity-matched comparison (§4.2). Dense early checkpointing (every 200 steps within epoch 1) locates an SFT checkpoint whose mean Δ_c matches GRPO within 0.2 points (2.6 vs. 2.5; $\approx$ 420 steps, GSM8K +4.1). Frozen FTS gives this checkpoint a suppression index of 49.3% [38.6, 59.8], compared with 38.7% at the matched-gain SFT checkpoint and 74.2% for GRPO at equal severity. The 24.9-point equal-severity gap has non-overlapping capability-clustered CIs; approximately 30% of the raw composition gap tracks damage magnitude and 70% tracks how the damage was produced. Together, the severity-matched and gain-matched comparisons isolate complementary axes and preserve the composition result.

Table 18: Scaling and family transfer at matched task gain. Suppression indices have capability-clustered 95% CIs; repair gain is the cost ratio to uniform replay at the common recovery target. Typing consistency uses externally verified subsets (36/39 for Qwen3-8B; n = 31–36 for other models). All rows use fixed FTS weights and the probe strategies in Appendix B.4.1.

Model	Suppr. index (RLVR)	Suppr. index (SFT)	Typing consist.	Repair gain
Qwen3-1.7B	62.4% [49.5, 74.6]	33.1% [22.2, 44.7]	89.2%	1.5×
Qwen3-4B	67.9% [55.8, 79.1]	36.5% [25.7, 47.9]	90.6%	1.7×
Qwen3-8B (main)	74.2% [63.4, 83.9]	38.7% [28.7, 49.1]	92.3%	1.8×
Qwen3-14B	78.6% [68.2, 87.6]	42.3% [30.9, 53.7]	90.1%	2.0×
Llama-3.1-8B-Inst.	69.8% [57.3, 81.1]	40.2% [28.6, 52.2]	88.7%	1.6×

Figure 5 plots these indices against parameter count, with Llama-3.1-8B-Instruct carried alongside the Qwen3 series as the cross-family point.

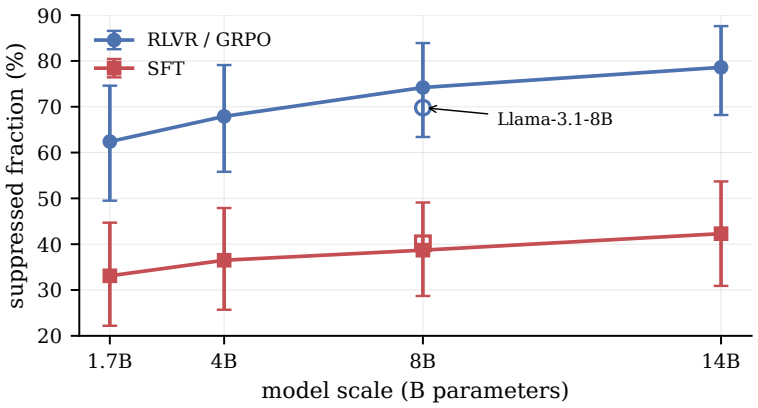

Figure 5: Scaling behavior of forgetting-type composition (§4.5).

Per-capability tables for all five models, all checkpoints, and both probe strategies (transferred vs. local) are provided in the released supplementary archive.

F.2 Category-Level MMLU Diagnosis Across Scales and Families

Category-level protocol. Each MMLU category uses 4.1k–7.6k anchors, at least 1.8k probe pairs, and its own reference curve on θ_0; median curve-fit $R^2 = 0.94$. The degradation gate is applied at benchmark level, and category losses are item-count weighted. Benchmark-level FTS closely matches the loss-weighted category mean (0.18 vs. 0.19 for SFT; 0.58 vs. 0.59 for RLVR). The category breakdown reveals actionable heterogeneity: routing with category-level FTS improves

MMLU recovery by 2.6 points at fixed budget. Appendix F.3 resolves the pattern further into individual disciplines.

Table 19: Fine-grained MMLU diagnosis by macro-category (Qwen3-8B, matched-gain checkpoints; §4.5). Suppression index is the FTS-weighted share of the category's loss mass; bootstrap 95% CIs. The bottom row repeats the benchmark-level §4.2 diagnosis.

	SFT (matched gain)			RLVR / GRPO (matched gain)		
MMLU category	Δ_c	FTS	**Suppr. frac.**	Δ_c	FTS	**Suppr. frac.**
STEM	7.1 ± 0.6	0.11	9.7% [5.3, 15.0]	2.6 ± 0.4	0.43	40.0% [29.5, 50.9]
Humanities	4.9 ± 0.5	0.27	21.6% [14.5, 29.7]	1.9 ± 0.3	0.69	65.6% [54.1, 75.8]
Social Science	5.2 ± 0.4	0.22	17.8% [11.3, 25.2]	2.0 ± 0.3	0.66	62.7% [51.0, 73.6]
Other (prof./misc.)	5.6 ± 0.7	0.19	15.2% [9.2, 22.4]	2.1 ± 0.4	0.62	58.3% [46.5, 69.7]
Benchmark-level (§4.2)	5.8 ± 0.4	0.18	—	2.1 ± 0.3	0.58	—

Cross-scale and cross-family replication. Category-level MMLU diagnosis is repeated on Qwen3-1.7B/4B/14B and Llama-3.1-8B-Instruct with frozen FTS v1.0 weights and each model's §4.5 probe-strategy branch (Appendix B.4.1); category-level reference curves are recomputed per model so the pair normalization of Eq. 2 remains within-category. Table 20 reports category-level FTS under both paradigms and the MMLU recovery gain of category-level over benchmark-level routing at fixed budget (closed-loop protocol, 2 seeds per model).

Table 20: Category-level MMLU FTS by model, paradigm, and macro-category, with the MMLU recovery gain (points) of category-level over benchmark-level FTS routing at fixed budget. The Qwen3-8B row restates Table 19. ST = STEM, Ot = Other, SS = Social Science, Hu = Humanities.

	SFT FTS				RLVR FTS				Routing
Model	ST	Ot	SS	Hu	ST	Ot	SS	Hu	**gain**
Qwen3-1.7B	0.08	0.15	0.18	0.23	0.33	0.51	0.54	0.61	+2.1
Qwen3-4B	0.09	0.17	0.21	0.24	0.38	0.56	0.60	0.66	+2.4
Qwen3-8B (main)	0.11	0.19	0.22	0.27	0.43	0.62	0.66	0.69	+2.6
Qwen3-14B	0.13	0.22	0.24	0.30	0.49	0.65	0.71	0.76	+2.3
Llama-3.1-8B-Inst.	0.10	0.18	0.21	0.26	0.41	0.59	0.62	0.67	+2.2

Cross-model pattern. The ordering STEM $<$ Other $<$ Social Science $<$ Humanities holds across all ten model–paradigm combinations. Category-level routing improves MMLU recovery by 2.1–2.6 points on every model, with no monotone trend in scale (Table 20).

F.3 Subject-Level MMLU Diagnosis (Ten Representative Disciplines)

To probe one level below the macro-categories of §4.5, we re-run the three-axis diagnosis on ten representative MMLU disciplines spanning all four categories (five STEM, three Humanities, one Social Science, one Other), on the same Qwen3-8B matched-gain checkpoints. Per-subject anchor pools (0.8k–2.1k examples) and probe pairs (≥ 400) are smaller than at category level, so bootstrap CIs are roughly 1.5–$2\times$ wider and per-subject curve fits retain median $R^2 = 0.91$; subject-level N_{50}^{ref} curves are recomputed on θ_0 so pair normalization stays within-subject. Table 21 reports point estimates.

Two regularities emerge. First, the coarse ordering is robust: every STEM subject sits below every Humanities subject in FTS under both paradigms, extending the category-level ordering of §4.5 to subject level. Second, the fine ordering is stable: rank-correlating the ten subjects' SFT and RLVR FTS gives Spearman $\rho = 0.94$, and the four discordant pairs out of 45 all involve subjects separated by less than one per-subject interval. Within-category variation is systematic and tracks formal adjacency to the mathematical training distribution: mathematics is the most overwritten subject overall (SFT FTS 0.06), biology lies at the suppressed end of STEM (0.16), and the two quantitative outliers of their categories—law within Humanities (0.62 vs. 0.71–0.74 under RLVR) and economics within Social Science (0.64 vs. category mean 0.66)—fall below their category means in the predicted direction. Loss-weighted subject means reproduce the category-level values of Table 19 within 0.02

Table 21: Subject-level MMLU diagnosis at matched gain (Qwen3-8B). FTS and loss-weighted suppression indices are point estimates; subject-level CIs are 1.5–2× wider than category-level CIs. Italic means cover all subjects in each category. Indices weight subject-level FTS by subject-level loss mass.

		SFT (matched gain)		**RLVR / GRPO (matched gain)**	
Subject	**Category**	FTS	**Suppr. frac.**	FTS	**Suppr. frac.**
Mathematics	STEM	0.06	5.1%	0.35	31.9%
Physics	STEM	0.09	7.8%	0.38	35.7%
Computer Science	STEM	0.10	9.4%	0.47	43.2%
Chemistry	STEM	0.13	11.2%	0.44	42.2%
Biology	STEM	0.16	13.9%	0.51	47.4%
STEM category mean		*0.11*	*9.7%*	*0.43*	*40.0%*
Law	Humanities	0.24	19.2%	0.62	58.9%
History	Humanities	0.28	22.7%	0.74	70.3%
Philosophy	Humanities	0.31	24.2%	0.71	68.2%
Humanities category mean		*0.27*	*21.6%*	*0.69*	*65.6%*
Economics	Social Science	0.20	16.4%	0.64	60.8%
Social Science category mean		*0.22*	*17.8%*	*0.66*	*62.7%*
Medicine	Other	0.22	17.4%	0.59	55.5%
Other category mean		*0.19*	*15.2%*	*0.62*	*58.3%*

FTS. The ten-discipline map already supports the routing decision of §4.5: allocate replay toward math-adjacent subjects and light operators elsewhere.

G Mechanistic Decomposition of the SFT–RLVR Gap

The analysis follows three steps: isolate trace source and proximity regularization in a matched factorial, relate the resulting recovery states to realized model–base distance, and trace the consequences for individual capabilities and repair cost.

G.1 Matched Factorial Design

Training arms. All arms use Qwen3-8B and stop within ± 0.5 GSM8K points of the main SFT arm's math gain.

- *SFT (§4.2 recipe)*: 210k external CoT, 2 epochs, LR 1×10^{-5}. Reference arm.
- *SFT + L2-SP*: identical data and schedule plus $\lambda\|\theta-\theta_0\|_2^2$, $\lambda = 10^{-3}$ selected as the largest value that still reaches matched gain within 3 epochs.
- *SFT + drift-matched*: LR 3×10^{-6}, 5 epochs, stopped at the checkpoint closest in $\|\delta\|$ to the KL arm (0.78 against 0.73; checkpoints are per epoch). This arm isolates update magnitude from explicit anchoring.
- *SFT + KL-to-base*: token-level forward KL to θ_0 on the training targets, $\beta = 0.01$—numerically the same penalty weight GRPO applies to its policy.
- *SFT, matched prompts/tokens* (†): teacher-written CoT restricted to the same 34k DeepScaleR prompts that survive rejection sampling, *passed through the same programmatic verifier used to filter the self-generated traces*, with up to 2 verified reference solutions retained per prompt so that the example count (58k) and total training tokens match the self-generated arm within 2%. The verifier rejects 4.1% of DeepScaleR reference solutions (answer-format mismatches and a residual of genuinely wrong derivations); the shortfall is topped up from the remaining verified solutions on the same prompts to hold the token budget.
- *SFT, matched prompts/tokens + KL* (†): the previous arm plus KL-to-base at $\beta = 0.01$.

- *SFT, self-generated* (†): for each of the 40k DeepScaleR prompts used by GRPO we sample 8 completions from θ_0 at temperature 1.0, keep the verifier-correct ones, and retain at most 2 per prompt (yielding 58k traces from 34k prompts with at least one success); LR 1×10^{-5}.
- *SFT, self-generated + KL* (†): the previous arm plus KL-to-base at $\beta = 0.01$.
- *GRPO,* $\beta = 0$: the §4.2 RLVR recipe with the reference KL term removed (gradient clipping and rollout settings unchanged).
- *GRPO,* $\beta = 0.01$: the §4.2 RLVR recipe.
- *PPO,* $\beta = 0.01$ (held-out prediction arm): PPO with clipped surrogate objective and a learned value head, on the same DeepScaleR prompts, the same programmatic verifiable reward, the same matched-gain stop, and the same nominal β as the GRPO recipe; 3 seeds. The arm is excluded from the factorial, ANOVA, and 13-arm regression and serves as the out-of-sample prediction check of §4.3.

Matched data construction. The four † arms vary trace authorship and KL regularization while sharing prompts, example count, token budget, and correctness filtering. Moving from the 210k external corpus to the matched 58k subset raises the suppression index from 38.7% to 40.5%. Both trace sources use the 34k prompts with verifier-correct base-model samples; the other 15% of prompts are excluded. Using unfiltered teacher traces gives 39.8%, compared with 40.5% after filtering, a 0.7-point effect alongside the 17.6-point matched teacher-to-self-generated contrast. Appendix J reports the prompt-difficulty comparison.

Table 22: Per-capability diagnoses for the factorial study (Qwen3-8B, fixed FTS, 3 seeds). The four † arms form the matched 2×2; the external-CoT reference appears in Table 1. Bottom rows report loss mass and the FTS-weighted suppression index.

	ext. + KL		**†matched**		**†match.+KL**		**†self-gen.**		**†self+KL**	
Capability	Δ_c	FTS	Δ_c	FTS	Δ_c	FTS	Δ_c	FTS	Δ_c	FTS
MMLU	4.8	0.32	5.5	0.21	4.8	0.33	3.3	0.42	2.6	0.52
MMLU-Pro	6.3	0.23	7.0	0.15	6.2	0.24	4.4	0.36	3.6	0.44
IFEval	6.2	0.64	8.2	0.45	6.1	0.65	7.0	0.56	3.8	0.77
AlpacaEval 2 LC	6.9	0.79	9.0	0.62	6.7	0.80	7.3	0.69	3.9	0.87
TruthfulQA	5.0	0.35	6.1	0.25	5.0	0.36	3.8	0.52	2.9	0.62
Safety	5.1	0.89	7.1	0.74	4.9	0.89	6.5	0.81	3.0	0.93
MGSM	4.0	0.33	4.8	0.22	3.7	0.35	3.4	0.46	2.6	0.58
Loss mass / mean	38.3	5.5	47.7	6.8	37.4	5.3	35.7	5.1	22.4	3.2
Suppression index	52.3%		40.5%		53.1%		58.1%		68.5%	

Factorial effects. The matched 2×2 analysis gives a $+16.5$-point effect for model-generated data ($F = 21.3$, $p = 0.0014$), $+11.5$ for proximity regularization ($F = 12.4$, $p = 0.006$), and a -2.2-point interaction ($F = 0.98$, $p = 0.34$). The combined 28.0-point shift reproduces 79% of the 35.5-point SFT–RLVR gap. Capability-clustered intervals are $[24.0, 31.9]$ for the shift, $[29.8, 41.1]$ for the gap, and $[73, 83]\%$ for their ratio, with $P(\text{ratio} > 0.6) > 0.99$.

Including the 1.8-point change from the external corpus to the matched subset brings the total shift to 29.8 points (84%). The remaining GRPO-to-self-generated-SFT+KL difference is 5.7 points (95% CI $[-3.1, 14.2]$, $p = 0.17$). At the unmatched recipe level, the corresponding main effects are $+17.8$ and $+12.0$ points. Statistical procedures are collected in Appendix J.

G.2 Realized Proximity Across Training Objectives

We sweep regularization strength in both objectives and compare the resulting models on a common measurement: KL to θ_0 on 2k held-out DeepScaleR prompts with eight samples each. This measures the realized effect of supervised regularization on training targets and RL regularization on rollouts. Table 23 also reports task-vector norm relative to the main SFT arm.

Pooled relation. Across the 13 matched-gain arms, measured KL predicts the suppression index with $R^2 = 0.962$, slope -50 points per unit KL, and Spearman $\rho = -0.99$. Adding trace source gives $R^2 = 0.979$ ($\Delta R^2 = 0.017$, $F(1, 10) = 8.1$, $p = 0.017$); adding objective identity gives

Table 23: Training arms and realized proximity to θ_0. KL is measured on held-out prompts; $\|\delta\|$ is relative to main SFT. The regression uses 13 matched-gain arms; PPO is the held-out prediction test. The SFT $\beta = 0.1$ arm reaches 82% of target gain and is reported separately.

Arm	β	KL	$\|\delta\|$	On-pol.	Suppr.	Matched?
SFT, external CoT	0	0.94	1.00	no	38.7%	yes
SFT, external CoT	10^{-3}	0.81	0.86	no	43.2%	yes
SFT, external + L2-SP	—	0.71	0.81	no	46.1%	yes
SFT, external, drift-matched	—	0.68	0.78	no	44.8%	yes
SFT, external CoT	10^{-2}	0.52	0.73	no	52.3%	yes
† SFT, matched pr./tok.	0	0.91	0.92	no	40.5%	yes
† SFT, matched + KL	10^{-2}	0.51	0.70	no	53.1%	yes
† SFT, self-generated	0	0.47	0.69	yes	58.1%	yes
† SFT, self-gen. + KL	10^{-2}	0.29	0.54	yes	68.5%	yes
GRPO	0	0.33	0.51	yes	66.9%	yes
GRPO	10^{-3}	0.28	0.44	yes	70.5%	yes
GRPO	10^{-2}	0.21	0.38	yes	74.2%	yes
GRPO	10^{-1}	0.11	0.29	yes	79.3%	yes
Held out of the fit: PPO as a second policy-gradient prediction test						
PPO, verifiable reward	10^{-2}	0.26	0.42	yes	71.8% (pred. 69.7%)	yes
Boundary arm: 82% of target gain, outside the matched-gain regression						
SFT, external CoT	10^{-1}	0.24	0.61	no	61.4%	no (82% of target)

$R^2 = 0.988$ ($\Delta R^2 = 0.008$, coefficient $+4.1$ points, $F(1, 9) = 6.2$, $p = 0.034$). Models that remain closer to the base leave more accessible losses across both training objectives.

Within-objective and resampling results. The nine supervised arms span KL 0.29–0.94 and give $R^2 = 0.936$ with slope -42; the four policy-gradient arms give $R^2 = 0.99$ with slope -55. Leaving out one arm at a time gives $R^2 \in [0.954, 0.971]$ and slopes in $[-52, -49]$. Measured KL has s.e. ≈ 0.02 against between-arm s.d. 0.27, giving reliability ratio 0.995 and attenuation-corrected slope -50.3.

Two views of proximity. Task-vector norm and measured KL both predict forgetting composition (Table 24): pooled R^2 is 0.973 and 0.962, respectively, with leave-one-out RMSE 2.55 and 3.06. Within supervised arms, KL gives $R^2 = 0.936$ and norm 0.919. The predictors correlate at Pearson $r = 0.979$ and Spearman 0.995; their joint model has variance-inflation factor 24.6. Together, they identify realized proximity as the common predictor across training objectives.

Table 24: Model comparison on the 13 matched-gain arms, response = FTS-weighted suppression index. ΔAIC is relative to the best single predictor (the task-vector norm); negative is better. LOO-RMSE is leave-one-arm-out prediction error in index points. The held-out PPO arm enters none of these fits.

Predictor(s)	R^2	ΔAIC	LOO-RMSE
$\|\delta\|$ (task-vector norm)	0.973	0	2.55
Measured KL to θ_0	0.962	$+4.2$	3.06
KL + $\|\delta\|$	0.978	-0.9	2.49
KL + on-policyness	0.979	-1.5	2.64
KL + on-policyness + objective label	0.988	-6.4	2.39
On-policyness alone	0.810	$+25.2$	6.89
Objective label alone	0.649	$+33.2$	9.07

Both distance measures predict composition more strongly than objective identity alone ($R^2 = 0.649$). The same nominal $\beta = 10^{-2}$ produces measured KL 0.52 under external-CoT SFT and 0.21 under GRPO. Realized distance therefore provides the useful link from a training configuration to the recovery state it leaves behind.

Table 25: Complete repair comparison (Qwen3-8B, 3 seeds). Each row averages the SFT- and RLVR-degraded scenarios. GPU·h columns itemize (a) diagnosis and operator search, (b) operator application, (c) replay, and (d) safeguards. Table 26 separates the two scenarios.

Method	**Total cost**	(a)	(b)	(c)	(d)	**Recovery**	**Retention**
Type-blind repair: no diagnosis of any kind							
LiNeS alone	0.6	–	0.6	–	–	46.3±2.9%	99.6%
Uniform replay (comparable cost)	42.0	–	–	39.1	2.9	71.2±2.6%	98.1%
Uniform replay (≈ 90% target)	79.6	–	–	73.8	5.8	89.4±1.8%	97.3%
Diagnosis ablations: routing machinery kept, type information destroyed							
Shuffled router (FTS permuted)	46.6	5.9	0.5	36.9	3.3	63.8±4.1%	98.3%
No-diagnosis router (FTS ≡ 0.5)	44.1	–	0.4	40.3	3.4	78.2±3.0%	98.4%
Lightweight-first and single-axis repair							
Try-light-first, then replay (greedy)	46.7	–	1.1	42.2	3.4	88.1±2.7%	98.5%
Axis-3-only router (LRR)	40.3	1.5	0.4	35.2	3.2	88.9±2.4%	98.7%
TRIAGE (full)	**44.2**	5.9	0.4	34.7	3.2	**91.7±1.5%**	**99.1%**
Oracle router (upper bound)	40.9	–	0.4	37.3	3.2	93.0±1.2%	99.2%

G.3 Capability Patterns and Repair Consequences

Per-capability pattern. The two levers do not act uniformly across capabilities. Knowledge and factuality—the capabilities SFT overwrites hardest—gain far more from on-policy data than from the anchor (MMLU FTS $0.21 \rightarrow 0.42$ under self-generation against 0.33 under KL alone; TruthfulQA $0.25 \rightarrow 0.52$ against 0.36), whereas chat, format, and safety inherit most of their benefit from the KL term (AlpacaEval $0.62 \rightarrow 0.80$ under KL against 0.69 under on-policy data; Safety $0.74 \rightarrow 0.89$ against 0.81). The crossover is mechanistically coherent: self-generated targets keep the token distribution inside the base model's support, so there is less pressure to rewrite factual circuitry, while the KL anchor restrains the alignment-layer drift that produces readout failures. It also shows up in the damage magnitudes—relative to the vanilla SFT vector, on-policy data removes 39–43% of the knowledge losses but only 17–24% of the chat and safety losses, and the KL term does the reverse—so the two factors cannot be summarized by a single severity scalar. Both effects are visible in the pre-fusion axes as well ($\tilde{N}_{50}$ falls by $2.8\times$ on knowledge between the extreme cells; G_c rises by 0.21 on chat).

Repair-cost consequence. Repairing the regularized self-generated SFT arm to the common target costs 43.4 GPU·h, compared with 71.3 for main SFT ($1.6\times$ cheaper). Replay's share of total repair cost falls from 84% to 61%. Self-generated traces and KL regularization thus improve both the loss profile and the cost of restoring it.

Cross-scale replication. On Qwen3-4B and 14B, the four matched corner arms (2 seeds, frozen weights) preserve the same ordering, and regularized on-policy SFT lands within 6.9 points of the corresponding GRPO arm (4B: 61.2% vs. 67.9%; 14B: 71.7% vs. 78.6%). The decomposition therefore replicates beyond the 8B configuration.

H Repair Evaluation and Cost Analysis

H.1 Repair Protocols and Cost Breakdown

Table 25 reports the complete policy comparison. Supplying oracle types to the same router gives 93.0% recovery, compared with 91.7% for FTS and a 94.2% linear-program upper bound. Algorithm 1 specifies the shared routing procedure.

Budget grid. Uniform replay swept over three total budgets ($\{0.35, 1.0, 3.0\}\times$ the TRIAGE realized budget); common-target numbers interpolate the resulting cost–recovery curve (Figure 6). All costs measured on 8×H800 wall-clock, converted to GPU·h.

The category split matches the expected boundary pattern: gains concentrate in suppression-dense categories (safety/format: up to $4.2\times$, where anchor data is also scarcest) and are modest in

Table 26: Per-scenario itemized repair costs (GPU·h) at the common $\approx 90\%$ mean-recovery target.

Scenario / method	(a) diag.	(b) light	(c) replay	(d) safegd.	**Total**	**Uniform**	**Gain**
RLVR-damaged, TRIAGE	5.9	0.5	9.5	1.2	17.1	41.5	2.4×
SFT-damaged, TRIAGE	5.9	0.3	59.9	5.2	71.3	117.7	1.7×
— safety capability only	0.9	0.1	2.1	0.4	3.5	14.6	4.2×
— knowledge capability only	0.9	0.1	31.8	1.9	34.7	44.1	1.3×

overwriting-dense knowledge capabilities (1.3×), where replay is genuinely necessary. Single-capability rows carry a proportionally smaller diagnosis charge (0.9 GPU·h for one capability instead of 5.9 for seven), which is why they are not simply the seven-capability figures divided down. **Data-count accounting**: at the common recovery target TRIAGE consumes 38k anchor *repair* examples vs. 96k for uniform replay (2.5×); this ratio counts repair data only and so runs ahead of the GPU·h ratio, which additionally carries the fixed diagnosis charge.

Additional baselines. The following use the same pooled-scenario convention as Table 25. *No repair* fixes the zero point: 0 cost, 0.0% recovery, 100.0% retention. *Task-vector interpolation* (WiSE-FT), the global-interpolation counterpart of the LiNeS row, costs 0.5—all of it light-operator time—for 41.8±3.4% recovery at 98.9% retention, so type-blind light repair stays below 50% recovery in both task-vector variants. The *Axis-1-only router* ($\tilde{N}_{50}$), the second single-axis ablation, costs 43.3 in total (3.9 diagnosis, 0.5 light operators, 35.6 replay, 3.3 safeguards) for 87.6±3.1% recovery at 98.8% retention; it is specified in Appendix H.1.1.

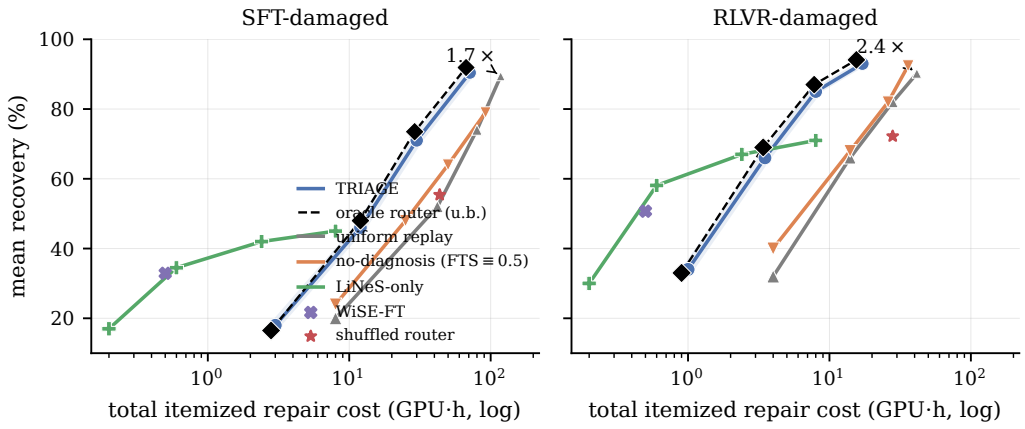

Figure 6: Repair cost–recovery Pareto frontiers (closed-loop; §4.4). TRIAGE reaches the common recovery target at 1.7× (SFT-damaged) to 2.4× (RLVR-damaged) lower total cost than uniform replay while retaining more of the new-task gain.

Safeguard sweep. Over KL weight $\{0.01, 0.1, 1.0\}$ × gate $\{95\%, 98\%, 99\%\}$: recovery varies within $[89.9, 92.4]\%$, total cost within $[42.4, 48.7]$ GPU·h, Pareto frontier displacement $\leq 11\%$. The 99% gate with KL 1.0 is the highest-retention corner (cost 48.7, retention 99.5%).

Constraint-strength sweep (ϵ_T). Re-running the full closed loop (diagnosis Axis 3 re-tuned, router re-executed) under three settings of the new-task retention constraint, averaged over the SFT- and RLVR-damaged scenarios (Qwen3-8B, 3 seeds), gives at $\epsilon_T = 0.5/1.0$ (default) /2.0 points: total cost 51.0/44.2/42.4 GPU·h, recovery 90.4±1.7/91.7±1.5/92.1±1.6%, retention 99.4/99.1/98.3%, and 1.6/1.8/1.9× gain over uniform replay. Tightening to 0.5 points disqualifies ∼14% of light-operator configurations—predominantly global task-vector interpolation with large α, whose retention cost scales with the interpolation step—so part of the suppressed repair load shifts to prompt-level operators and small targeted replay, raising cost while improving retention. Loosening to 2.0 points changes the solution little, since light operators sit far from the constraint

boundary at the optimum. The routing advantage over uniform replay at the common recovery target persists at every setting.

H.1.1 Single-Axis and Lightweight-First Baselines

The three ablated routers—two in Table 25, the Axis-1-only variant above—are specified as follows. *Axis-3-only*: run Axis 3 alone (1.5 GPU·h), send capabilities with $\mathrm{LRR}_c \geq \eta = 0.8$ to their best light operator, and allocate replay to the rest in proportion to $(1 - \mathrm{LRR}_c)\Delta_c$—i.e. the full router with LRR substituted for FTS. *Axis-1-only*: identical but keyed on the frozen $\tilde{N}_{50} = 0.5$ threshold, at 3.9 GPU·h of diagnosis. *Try-light-first*: no diagnosis; for each degraded capability apply every light operator in a fixed order, accept the first that clears η on the validation slice, and otherwise fall back to replay with budget proportional to Δ_c alone. The greedy baseline is the most practitioner-realistic alternative to TRIAGE and its cost premium comes from two sources measured separately: 0.7 GPU·h of light-operator attempts that end up discarded, and 7.5 GPU·h of excess replay caused by allocating on Δ_c instead of on $(1 - \mathrm{FTS}_c)\Delta_c$.

Table 27: Where fusion matters: recovery at comparable cost for the Axis-3-only router vs. full TRIAGE, across settings. The gap widens off the calibration family and in the small-Δ regime where LRR is least reliable.

Setting	Axis-3-only	TRIAGE (full)	Gap
Qwen3-8B, SFT- and RLVR-damaged (main)	88.9±2.4%	91.7±1.5%	2.8
Qwen3-8B, RLVR-damaged only (small Δ_c)	86.1±3.3%	91.3±1.8%	5.2
Llama-3.1-8B-Instruct (cross-family)	82.4±3.8%	87.1±2.2%	4.7
Qwen3-1.7B (cross-scale)	85.7±3.1%	89.2±2.0%	3.5
Mixed-band instances only ($\mathrm{FTS} \in [0.3, 0.7]$)	79.8±4.2%	86.6±2.6%	6.8

Matched-recovery comparison. At the measured marginal replay rate, extending Axis-3-only repair to 91.7% recovery costs an estimated additional 5.3 GPU·h, or 45.6 in total, compared with 44.2 for TRIAGE. The corresponding estimate for try-light-first is 53.5 GPU·h.

Where fusion contributes most. The margin is largest when direct operator success gives an incomplete allocation signal: small degradations (+5.2 points), cross-family repair (+4.7), and mixed-band losses (+6.8). Recovery dynamics and probe readability add information about replay demand and retained accessibility, sharpening the budget decisions in these regimes (Table 27).

Replay-allocation sensitivity. On ten real-checkpoint instances spanning FTS 0.12–0.83, including four mixed cases, the linear allocation predictor $(1-\mathrm{FTS}_c)\Delta_c$ gives $R^2 = 0.92$ for realized recovery per unit budget. Adding a quadratic term gives $p = 0.23$ in a likelihood-ratio test. Using exponents 1.2 on either $(1 - \mathrm{FTS}_c)$ or Δ_c changes total cost by less than 3% and recovery by less than 0.5 points in the full repair loop.

H.2 Repair Compatibility and Diagnosis Perturbations

This section expands the repair-compatibility result in §4.4.

2×2 cross-application. Applying each repair family to each diagnosed type (Table 4) reveals a clear compatibility pattern. Lightweight operators recover 88.4% of suppression-like losses but only 13.7% of overwrite-like losses under the tested family. Large-budget replay also recovers suppression-like losses (90.2%), but costs 7.4× more per recovered point and reduces the new-task score by 2.1 points. The value of diagnosis is therefore primarily allocation: it avoids expensive replay where a lightweight repair is already effective.

Diagnosis perturbations. Flipping 20% of type assignments before routing reduces recovery by 11.8 points and raises cost by 38%, with a local slope of ≈ 0.6 recovery points per 1% diagnostic error (Figure 7). Observed disagreements concentrate in the mixed band, where FTS changes the replay budget continuously.

Fusion-margin decomposition. The 3.8-point typing-accuracy gain over Axis 3 (92.3% vs. 88.5%), combined with the measured flip slope, accounts for 2.3 recovery points. Disabling the mixed band

while retaining fused scores reduces closed-loop recovery by another 0.5 points. Together, the two components account for the observed 2.8-point fusion margin.

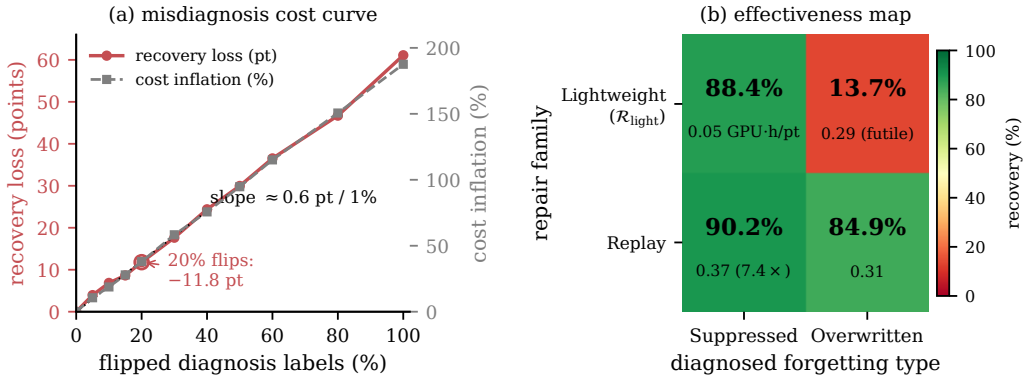

Figure 7: Diagnosis perturbations: routing benefits collapse under corrupted diagnosis; the operator × type boundary map delineates where each repair technique works.

I Transfer to a Second Training Objective

I.1 Diagnosis Under Instruction-Following Post-Training

We extend the diagnosis from mathematical-reasoning post-training to a second objective: *verifiable instruction following* on Qwen3-8B. The evaluation follows the matched-gain protocol with fixed FTS v1.0 weights, supported by the 24 instruction-objective injections described below. SFT trains on a 140k Tulu-style instruction mixture; GRPO uses verifiable constraint rewards (IFEval-style programmatic checks) on instruction prompts; both runs stop at matched IFEval gain (+8.7 SFT vs. +9.1 GRPO, within 0.4 points). Because instruction following and open-ended chat are now the training target, the diagnosed capability set replaces IFEval/AlpacaEval with mathematical reasoning (GSM8K, MATH500), retaining seven capabilities. Diagnosis uses transferred probes and 3 seeds; anchor pools follow the §B.3 construction with math pools drawn from NuminaMath validation-disjoint splits.

Table 28: Cross-objective forgetting anatomy: instruction-following post-training on Qwen3-8B at matched IFEval gain. Same conventions as Table 1; The v1.0 score is applied without recalibration; results use 3 seeds.

	Instruction SFT			**Instruction GRPO (RLVR)**		
Capability	Δ_c	FTS	**T**	Δ_c	FTS	**T**
MMLU (knowledge)	3.4±0.4	0.52	M	2.1±0.3	0.68	M
MMLU-Pro	4.1±0.5	0.44	M	2.4±0.3	0.62	M
GSM8K (math)	6.2±0.6	0.26	O	3.0±0.4	0.62	M
MATH500	5.4±0.7	0.22	O	2.7±0.4	0.56	M
TruthfulQA	2.9±0.4	0.64	M	2.0±0.3	0.74	S
Safety (AdvBench+XSTest)	4.6±0.6	0.82	S	2.5±0.3	0.88	S
MGSM (multiling.)	3.8±0.5	0.49	M	2.3±0.3	0.57	M
Mean degradation	4.3 points			2.4 points		
FTS-weighted suppression index	**45.6**% [35.9, 55.4]			**66.3**% [56.7, 75.7]		

Three observations follow from Table 28. *First, the paradigm-level ordering replicates*: the suppression index under RLVR (66.3%, capability-clustered 95% CI [56.7, 75.7]) exceeds that under SFT (45.6% [35.9, 55.4]), with non-overlapping intervals. *Second, the effect adapts to the training distribution*: the gap is 20.7 points here versus 35.5 on the math objective (Table 1), and the most-overwritten capabilities shift to mathematical reasoning (FTS 0.22–0.26), complementing the

STEM-adjacent pattern under math training (§4.5). Thus, the detailed composition follows the task distribution while the paradigm ordering persists. *Third, diagnosis transfers with frozen weights*: the 24 instruction-objective injections yield AUROC 0.905; a 6-instance extreme-FTS check passes $6/6$; and the closed loop reproduces both the repair advantage ($1.9\times$) and fusion margin ($+2.9$ points) measured on math. An independent refit keeps every weight inside its v1.0 bootstrap interval.

Independent cross-task recalibration. We refit the fusion weights on 24 mechanism-matched instruction-objective injections (the Table 8 recipes re-targeted at instruction-following and chat capabilities, two strengths, plus 4 replicates on the highest-seed-variance cells, Qwen3-8B), obtaining $(w_1', w_2', w_3', b') = (1.79, 1.04, 1.31, -0.18)$. Relative to frozen v1.0 $(1.84, 0.96, 1.37)$, the drift is $\Delta w = (-0.05, +0.08, -0.06)$, with every weight inside its v1.0 bootstrap CI (Appendix D.1). The refit and frozen rankings correlate at Spearman $\rho = 0.92$, above the 0.90 criterion, and preserve $w_1 > w_3 > w_2$ in 86% of bootstrap resamples. The small drift quantifies objective-dependent variation while preserving the ordering consumed by the router.

I.2 Repair Under Instruction-Following Post-Training

We evaluate the diagnosis-to-repair link on the instruction-following checkpoints of Appendix I.1 with the four routers that carry the central comparison: one type-blind baseline, two lightweight diagnosis routers, and full TRIAGE. This compact transfer experiment preserves the common recovery-target protocol and isolates the method ordering and fusion margin on a second objective.

Protocol: Qwen3-8B, both instruction-objective damage scenarios (SFT- and GRPO-induced) pooled, 3 seeds, seven capabilities, frozen FTS v1.0 weights and transferred probes, $\mathcal{R}_{\text{light}}^{\text{v1.0}}$ v1.0 unchanged, and the same Algorithm 1 thresholds. Uniform replay is run to the same approximately 90% recovery target so the cost ratio is comparable to Table 25.

Table 29: Cross-objective reduced closed loop on the instruction-following objective (Qwen3-8B, SFT- and GRPO-damaged scenarios pooled, 3 seeds). Cost in GPU·h, itemized as in Table 25: (a) diagnosis, (b) light operators, (c) replay, (d) safeguards.

Method	**Total**	(a)	(b)	(c)	(d)	**Recovery**	**Retention**
Uniform replay ($\approx 90\%$ target)	83.4	–	–	76.7	6.7	$89.8\pm2.0\%$	97.1%
Try-light-first, then replay	44.1	–	1.0	39.8	3.3	$86.4\pm3.0\%$	98.3%
Axis-3-only router	38.7	1.5	0.4	33.6	3.2	$87.2\pm2.7\%$	98.5%
TRIAGE (full)	**43.0**	5.9	0.4	33.5	3.2	$\mathbf{90.1\pm1.8\%}$	$\mathbf{98.9\%}$

The instruction-following loop reaches 90.1% recovery at 43.0 GPU·h, a $1.9\times$ cost advantage over uniform replay. The four methods retain their ordering, with a 2.9-point fusion margin over Axis-3-only repair. Thus, both the allocation benefit and the value of multi-axis diagnosis transfer to the second objective.

J Statistical Analysis and Precision

Bootstrap resampling units. Suppression-index CIs in the main text are *capability-clustered*: we resample the seven capabilities with replacement and, within each drawn capability, resample its three seeds, matching the population of capabilities over which the result generalizes. Instance-level bootstrap gives narrower intervals ($[30.9, 46.8]$ for SFT and $[66.1, 81.5]$ for RLVR); the reported clustered intervals use seven capabilities as the effective sample size. For per-capability quantities (Tables 19, 11), the resampling unit is the evaluation item, as stated in each caption.

J.1 Composition Indices and Resampling

Continuous and discrete summaries. The FTS-weighted suppression index is $\text{index} = \sum_c w_c \Delta_c / \sum_c \Delta_c$, with $w_c = \text{FTS}_c$: each capability contributes its lost performance weighted by continuous accessibility. Table 30 also reports discrete capability counts and two alternative mixed-band assignments. Convention S sets mixed-band weights to 1; convention O sets them to 0; other weights remain at FTS.

Table 30: Convention sensitivity of the §4.2 headline (Qwen3-8B, matched gain). Mixed-band mass is the share of total degraded points carried by capabilities with $\text{FTS} \in [0.3, 0.7]$.

Convention for mixed-band mass	SFT	RLVR	Gap	Ordering
Proportional to FTS (reported)	38.7%	74.2%	35.5	RLVR > SFT
All to suppressed pole (convention S)	56.6%	95.3%	38.7	RLVR > SFT
All to overwritten pole (convention O)	19.8%	44.7%	24.9	RLVR > SFT
Mixed-band mass	36.8%	50.6%	—	—
Discrete capability count (S : M : O)	1:2:4	3:4:0	—	7/7 paired, $p = 0.016$

The RLVR–SFT gap ranges from 24.9 to 38.7 points across these assignments. The discrete summary gives three suppression-dominant and zero overwrite-dominant capabilities for RLVR, versus one and four for SFT. All seven paired FTS differences are positive ($+0.22$ to $+0.44$; exact sign-test $p = 0.016$).

J.2 Measurement Precision and Degradation Thresholds

Precision by loss magnitude. We re-evaluate each axis input ten times, resampling evaluation, decoding, and judge seeds. For losses $\Delta_c \in [2, 4)$ ($n = 24$), $[4, 8)$ ($n = 21$), and ≥ 8 ($n = 13$), FTS standard deviations are 0.041, 0.019, and 0.014, and ranking agreements are Kendall-$\tau = 0.94$, 0.97, and 0.98. Three label changes occur, all inside the mixed band of the smallest-loss stratum.

Monte Carlo propagation gives evaluator-noise contributions of ± 1.8 points to the RLVR suppression index and ± 0.9 to SFT. Replacing the recovery denominator by $\Delta_c + 0.5$ gives indices 72.9% and 38.1%. The equal-severity comparison in Appendix F.1 gives 49.3% for early SFT, a 24.9-point gap from GRPO.

Degradation threshold. The gate $\Delta_c \geq \tau_{\min} = 2$ applies to the three-seed mean. All seven capabilities qualify under both paradigms; the smallest means are 2.0 for RLVR and 4.7 for SFT. Of 42 individual capability–paradigm–seed cells, two RLVR cells fall below the gate (MGSM 1.8, MMLU 1.9); the smallest SFT cell is 4.2. Bootstrap resampling includes all three seeds of each qualifying capability. A per-seed gate gives an RLVR index of 74.0%, compared with 74.2% under the mean gate; excluding the two affected capabilities gives 79.1%.

J.3 Factorial Inference and Prompt Difficulty

Factorial inference. The two-way analysis treats capability as a random effect. Main effects use their capability-interaction mean squares with Satterthwaite denominators. The reproduced-gap ratio is computed jointly within each capability-clustered bootstrap sample.

Prompt difficulty. At fixed token budget, teacher-CoT training on the 34k base-solvable prompts gives a 40.5% suppression index. Sampling from the full prompt set while matching the subset's

mean pass rate gives 39.6%, a 0.9-point change compared with the 17.6-point matched teacher-to-self-generated contrast.