THE ARITHMETIC OF LLM POST-TRAINING: MERGING AND DISTILLATION ON A MOVING RECIPIENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are post-trained through successive stages such as SFT, RLVR, and OPD, yet it remains unclear *how an update learned along one training path composes with a model that has already progressed along another*. We study this question on Qwen3 models from 0.6B to 8B parameters by fixing an update learned in one training lineage (the donor) and carrying it, by merging or by distillation, to successive checkpoints of a second lineage (the recipient) that differs only in its random seed. The experiments show three regularities in how a donor update composes with its recipient. (1) Post-training updates are portable: a transferred update preserves the signature of its teacher, the problems it gains and the problems it loses. (2) Transfer is recipient-dependent: the same update improves an early recipient and harms a mature one, whether it was learned by reinforcement learning or by distillation. (3) Averaging and distillation diverge as the recipient matures: weight averaging leads from the stage where direct addition stops transferring, because it preserves more of the recipient's existing capability. A decomposition into problems that only the teacher solves and problems that only the recipient solves explains the second and third regularities: as the recipient advances, there is less to acquire and more to lose, and distillation keeps giving up the recipient's own problems while averaging retains them. These results show that the effect of a post-training update depends jointly on what the update encodes and what the recipient has already learned. The code is available here.

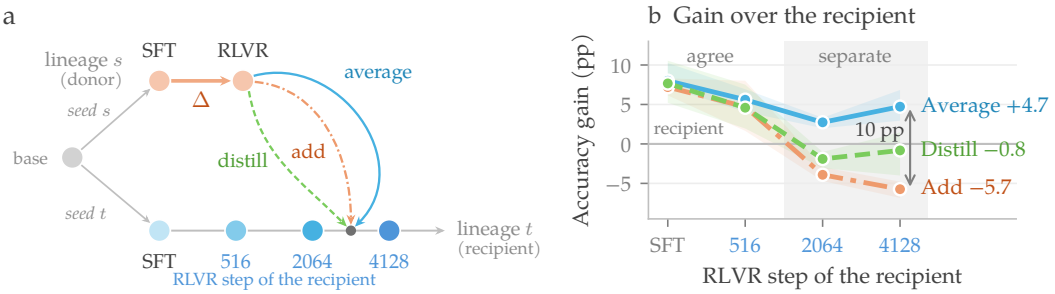


Figure 1: The update that RLVR leaves in one lineage is carried to another lineage of the same pipeline while that lineage is still training. (a) Two 0.6B lineages start from one base checkpoint and differ only in their training seed. The update Δ of lineage s is added to, averaged with, or distilled from its RLVR model into checkpoints of lineage t saved along its own RLVR stage. (b) The accuracy gains of the three operations over the checkpoint they act on agree on early checkpoints and separate later. Addition then loses and distillation falls to near zero, while averaging keeps a gain that ends 10 points above addition. Lines are means over directed pairs and evaluation sets, and bands show their range.

1 INTRODUCTION

Large language models (LLMs) are post-trained through a sequence of stages: supervised fine-tuning (SFT), reinforcement learning with verifiable rewards (RLVR), and on-policy distillation (OPD) (Guo et al., 2025; Yang et al., 2025). Each stage leaves a parameter update, the difference

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.

between the checkpoint it produces and the checkpoint it starts from, and current practice reuses these updates in three ways. Model merging adds, scales, and averages fine-tuning differences as task vectors (Ilharco et al., 2023; Yadav et al., 2023; Wortsman et al., 2022). OPD distills from a separately trained teacher whose capability the student is expected to acquire on top of its own (Agarwal et al., 2024; Yang et al., 2025). Fusing a checkpoint back toward its base reuses part of an earlier update (Liu et al., 2026). We call these operations, together with the rules that determine when their result equals the sum of its parts, *the arithmetic of post-training*.

Merging assumes that updates obey the rules of vector addition wherever they are added. Work on the SFT-to-RL handoff suggests otherwise. The state one stage leaves shapes the next, and an over-trained SFT checkpoint loses plasticity toward RL (Liu et al., 2026; Kang et al., 2026). The rules of the arithmetic may therefore depend on how far the receiving model has progressed in its own training. We ask what arithmetic merging and distillation obey on a moving recipient, a checkpoint taken at successive points of its own RLVR stage.

Measuring these rules requires a second, independently trained lineage. Within one lineage the RLVR checkpoint equals the SFT checkpoint plus the update, so adding the update back measures nothing. The update has to stay fixed while the recipient moves, and only a recipient from another lineage moves independently of it. The settings of practical interest are also cross-lineage, often with a recipient that is still training. Programs merge the checkpoints of parallel lineages (Wortsman et al., 2022), and pipelines distill a student at an arbitrary point of its RLVR stage from a teacher trained as another lineage of the same program (Yang et al., 2025). Our lineages start from one base, train on one data order under one protocol, and differ only in their seed (Figure 1a), so a failure to transfer reflects the recipient's state.

> ❶ *What kind of object is the update that RLVR or OPD leaves behind, and does adding it to another checkpoint transfer its effect?*
> ❷ *Where does addition cease to transfer, and does the answer depend on the update's origin?*
> ❸ *When two updates meet on one checkpoint, does averaging them compose what distillation between their owners does not?*

The three questions share one first-order account, the expansion that underlies the analysis of task arithmetic (Ortiz-Jimenez et al., 2023). For a checkpoint θ with expected accuracy $J(\theta)$ and an update Δ formed on another checkpoint, the gain of adding $\alpha\Delta$ is

$$J(\theta + \alpha\Delta) - J(\theta) = \underbrace{\alpha \langle \nabla J(\theta), \Delta \rangle}_{\text{first-order transfer}} + \underbrace{O(\alpha^2)}_{\text{response of the recipient}} . \tag{1}$$

The first term transfers the update's effect, and the second is the recipient's response to being moved. We use the two terms to organize the results. The update carries its teacher's identity and transfers through the first term to checkpoints close to the one that formed it. The first term shrinks as the recipient advances along its own stage, and at full dose the response term dominates.

The results suggest a diagnostic for pipelines. Before choosing an operation, a pipeline can count the recipient-only problems, solved by the recipient but not the teacher, and the teacher-only problems. At 0.6B these counts mark the recipients on which addition loses and averaging leads distillation.

❶ *Portable updates.* The update that OPD leaves is a task vector that carries its teacher's identity: on an early checkpoint of another lineage it gains and loses the problems its teacher makes students gain and lose. The shared lost problems rule out teacher strength as the explanation, and a sign shuffle and a negation show that the direction carries the effect (Section 4.1).

❷ *Limits of addition.* At 0.6B the same RLVR update gains less at every checkpoint along the recipient's own RLVR trajectory and loses at the mature one, and the OPD update changes sign in the same way. Small doses rule out an overshoot. Extending the recipient's own update by half the donor's norm costs 3 to 11 points, where half of the donor changes accuracy by about one point or less. On the 1.7B and 8B pairs the benefit of a donor falls when the recipient keeps its full own update instead of half of it, in all thirty-two comparisons, thirty with intervals excluding zero (Sections 4.2 and 5.1).

❸ *Averaging versus distillation.* At 0.6B, averaging and distillation coincide on most early recipients and separate in the same interval of the stage in which addition changes sign; at 1.7B

and 8B the average leads on point estimates from the first RLVR checkpoint on. At 0.6B, once addition turns negative, averaging leads by 3 to 8 points, partly because it retains more of the problems only the recipient solves. Distillation from the peer changes total accuracy by a few points relative to distillation from the recipient's own checkpoint, yet on mature recipients it raises accuracy on the teacher-only problems by up to 34 points and lowers it on the recipient-only ones by up to 35. Controls rule out a retreat toward SFT, the token signal's sign, and the merging operator. Distillation from a larger teacher forfeits less and matches or beats averaging at every RLVR checkpoint of the recipient (Sections 4.3 and 5.2).

2 Related Work

Interaction of post-training stages. Studies of the SFT-to-RL handoff find that memorization and generalization split between the stages (Chu et al., 2025; Shenfeld et al., 2026), that a high SFT score can predict a poor RL outcome (Kang et al., 2026), that the order of OPD and RLVR changes the winner (Li et al., 2026a), and that an over-trained SFT checkpoint loses plasticity toward RL (Liu et al., 2026; Dohare et al., 2024). We ask how the recipient's state shapes an update from a peer.

On-policy distillation. On-policy distillation trains the student on its own samples under token-level teacher supervision (Agarwal et al., 2024; Gu et al., 2024; Song & Zheng, 2026), and industrial pipelines apply it after RLVR with teachers trained in the same program (Yang et al., 2025). Analyses of the stage vary the teacher or the signal (Li et al., 2026b;c; Ding & Zhang, 2026; Wang et al., 2026b; Feng et al., 2026; Yu et al., 2026a); we fix the teacher and vary the student's stage.

Model merging and task arithmetic. Model merging adds and subtracts fine-tuning differences as task vectors (Ilharco et al., 2023), trims their interference (Yadav et al., 2023; Yu et al., 2024), finds that earlier task vectors merge better (Zhou et al., 2026), and averages models fine-tuned from one start with gains that decay in the number merged (Wortsman et al., 2022; Jang et al., 2024; Frankle et al., 2020; Wang et al., 2026a). RLVR and OPD updates are sparse (Mukherjee et al., 2025; Balashov, 2025; Shen et al., 2026; Yu et al., 2026b) yet can merge poorly (Wu et al., 2026), and an ensemble combines the features its members learn separately (Allen-Zhu & Li, 2023). These works fix the recipient and vary the update; we fix the update and move the recipient.

3 Study Design

3.1 Peer Lineages and a Moving Recipient

Answering these questions needs two training lineages that are peers in every respect except their own progress. In the main design every model is Qwen3-0.6B (Yang et al., 2025), and lineages $r \in \{1, 2, 3\}$ start from its base checkpoint, each trained by SFT and then RLVR on the same data under the same protocol and differing only in the random seed. The OPD jobs of the student-by-teacher matrix start from the SFT checkpoints. Within lineage r, S_r is the SFT checkpoint and $R_r[K]$ is the RLVR checkpoint saved after K of 4,128 optimizer updates, with $K \in \{516, 1032, 2064, 4128\}$. The mature checkpoint $R_r = R_r[4128]$ is the end of the stage (Figure 1a), and we write $R_r[0] = S_r$. Lineage 1 scores lower at its mature checkpoint than at $K = 2064$, while lineage 2 still improves on the in-distribution set up to $K = 4128$.

Each update is the difference between two checkpoints of one lineage. The RLVR update of lineage s is $\Delta_s^{\mathrm{R}} = R_s - S_s$, and the OPD update $\Delta_{r,s}^{\mathrm{O}} = O_{r\to s} - S_r$ is the endpoint $O_{r\to s}$ of OPD from S_r under the frozen teacher R_s, minus its start. Distilling each of the three SFT checkpoints under each of the three mature RLVR teachers gives a student-by-teacher matrix of nine OPD jobs, and the update of a job can be ported to a lineage that supplied neither its student nor its teacher. Each directed pair (t, s) of lineages, written $t \leftarrow s$, assigns the roles. Lineage s supplies a fixed donor update and its mature checkpoint as teacher, and lineage t supplies the recipient at successive stages K of its own RLVR training. Figures label each pair by its recipient lineage. Since the lineages share base, data, and protocol, a change in the update's effect along K reflects the recipient's progress.

The stage experiments at 0.6B use lineages 1 and 2 as one pair in both directions, and the third lineage enters the student-by-teacher matrix and the six-pair contrasts. Pairs of lineages at 1.7B

and 8B repeat the design, and a 4B pair serves as a boundary case (Appendix H). The mature 1.7B checkpoints also serve as larger teachers for the 0.6B recipients. SFT uses 14,448 problems from OpenR1-Math-220k (Hugging Face, 2025) with one verifier-correct response each. RLVR uses GRPO with a verifier reward (Shao et al., 2024; Yu et al., 2025), and OPD uses the sampled-token form of on-policy distillation (Agarwal et al., 2024). Both train on a disjoint set of 14,448 mathematics prompts (Yu et al., 2025). All jobs use verl (Sheng et al., 2025) with reasoning mode off. The lineages, the matrix, and the six-pair distillation jobs of Table 11 use a first protocol with two passes. The comparison of averaging with distillation uses a second protocol with a larger prompt batch, one pass, and a corrected OPD loss normalization (Appendix A). Every checkpoint is evaluated by its greedy accuracy Y, in percentage points, on $\mathcal{E}_{\text{in}}$, 1,024 problems drawn from the distribution of the training prompts but disjoint from them, and on $\mathcal{E}_{\text{ext}}$, 2,960 external competition problems. At 0.6B the SFT checkpoints solve about 7% of $\mathcal{E}_{\text{in}}$ and the mature checkpoints 10% to 17%. Each comparison pairs two checkpoints problem by problem, with 95% paired bootstrap intervals over problems. Each cell is one comparison for one directed pair of lineages on one evaluation set, so counts such as "six of eight cells" report how many intervals exclude zero.

3.2 Carrying an Update to the Recipient

Addition, averaging, TIES merging, and distillation carry the update of lineage s to lineage t.

Definition 1 (Recipient, donor, operations)**.** The recipient at stage K of lineage t is the checkpoint $\theta = S_t + u_t(K)$ with own update $u_t(K) = R_t[K] - S_t$. The donor is an update Δ formed on another lineage, with teacher $T = R_s$, the donor lineage's mature checkpoint. Operations map θ to θ'. Addition at dose α gives $\theta + \alpha\Delta$, and averaging gives $A_K = S_t + \frac{1}{2}u_t(K) + \frac{1}{2}\Delta$. Distillation gives the endpoint of one pass of sampled-token OPD from θ under the frozen teacher policy π_T with per-token advantage

$$\hat{a}_j(x, y) = \log \pi_T(y_j \mid x, y_{<j}) - \log \pi_\theta(y_j \mid x, y_{<j}) \tag{2}$$

on each token y_j of responses $y \sim \pi_\theta(\cdot \mid x)$. The effect of an operation is its change of accuracy

$$g = Y(\theta') - Y(\theta). \tag{3}$$

The average is the midpoint between the recipient and $S_t + \Delta$, and TIES merging (Yadav et al., 2023) replaces addition with a trimmed, sign-aligned merge.

Each operation comes with controls that remove one alternative explanation each (Table 3). The direction is tested by a sign shuffle that flips the sign of each coordinate at random, a projection that keeps, per weight matrix, the singular directions that carry 90% of the squared singular values, and the negation of Δ. Doses $\alpha \in \{0.25, 0.5\}$ test for an overshoot. The own-direction extension, which lengthens the recipient's own update by the norms of 0.25Δ and 0.5Δ, tests for a generic cost of moving. The shortened own update $H_K = S_t + \frac{1}{2}u_t(K)$ tests for a retreat toward SFT. Distillation under the recipient's frozen checkpoint, the self-teacher, measures the drift of training. The effect of the donor's teacher, the cross teacher, is its distillation endpoint minus that of the self-teacher.

3.3 Reading a Change Problem by Problem

The total accuracy change hides which problems moved, so we split problems by who solves them.

Definition 2 (Strata)**.** Let $c_\theta(x) \in \{0, 1\}$ be the greedy correctness of θ on problem x of an evaluation set $\mathcal{X}$ of size N, so that $Y(\theta) = \frac{100}{N}\sum_x c_\theta(x)$. The pair (θ, T) partitions $\mathcal{X}$ into $\mathcal{U}_T = \{x : c_T(x) = 1, c_\theta(x) = 0\}$, $\mathcal{U}_R = \{x : c_T(x) = 0, c_\theta(x) = 1\}$, $\mathcal{U}_B = \{x : c_T(x) = c_\theta(x) = 1\}$, and $\mathcal{U}_N = \{x : c_T(x) = c_\theta(x) = 0\}$, with weights $w_\sigma(\theta) = |\mathcal{U}_\sigma|/N$ and, for the output θ' of an operation, rates $r_\sigma = \frac{100}{|\mathcal{U}_\sigma|}\sum_{x \in \mathcal{U}_\sigma}\big(c_{\theta'}(x) - c_\theta(x)\big)$.

Proposition 1 (Stratum identity and its stage split; proof in Appendix C)**.** *For every operation,*

$$g = \sum_{\sigma \in \{T,R,B,N\}} w_\sigma(\theta)\, r_\sigma, \tag{4}$$

and between two stages with weights w_σ, w'_σ and rates r_σ, r'_σ,

$$g' - g = \underbrace{\sum_\sigma (w'_\sigma - w_\sigma)\, r_\sigma}_{\textit{weights move}} + \underbrace{\sum_\sigma w_\sigma\,(r'_\sigma - r_\sigma)}_{\textit{rates move}} + \sum_\sigma (w'_\sigma - w_\sigma)(r'_\sigma - r_\sigma). \tag{5}$$

We call $w_\sigma r_\sigma$ the share of stratum σ in g. The teacher-only problems $\mathcal{U}_T$ contain everything an operation can acquire from the teacher, and the recipient-only problems $\mathcal{U}_R$ contain everything it can forfeit. The weights describe the recipient's state: from the first RLVR checkpoint to the mature one, $\mathcal{U}_R$ grows and $\mathcal{U}_T$ shrinks. We say an operation exchanges strata when it acquires teacher-only problems while forfeiting recipient-only ones. The rates describe the operation: r_T is its acquisition rate and $-r_R$ its forfeiture rate. The acquisition rate of a fixed donor measures the recipient's plasticity toward it. The stage analyses define the strata at each stage from $R_t[K]$ and R_s.

3.4 Why Addition Should Fade

Proposition 2 (First-order gain of addition)**.** *Let $v(x, y) \in \{0, 1\}$ be the verifier reward, $p_x(\theta) = \mathrm{E}_{y\sim\pi_\theta(\cdot|x)}[v(x, y)]$, and $J(\theta) = \mathrm{E}_x[p_x(\theta)]$ the RLVR objective, of which Y is the greedy reading. If π_θ is twice differentiable in θ, then for every α there is a $\tilde{\theta}$ on the segment from θ to $\theta + \alpha\Delta$ with $J(\theta + \alpha\Delta) - J(\theta) = \alpha\langle\nabla J(\theta), \Delta\rangle + \frac{\alpha^2}{2}\Delta^\top\nabla^2 J(\tilde{\theta})\Delta$, and with the Fisher matrix $F_x(\theta) = \mathrm{E}_y[\nabla \log \pi_\theta \nabla \log \pi_\theta^\top]$ every problem satisfies*

$$|\langle\nabla p_x(\theta), \Delta\rangle| \le \sqrt{p_x(\theta)\,(1 - p_x(\theta))}\,\sqrt{\Delta^\top F_x(\theta)\,\Delta}. \tag{6}$$

The expansion is accurate only close to the expansion point (Ortiz-Jimenez et al., 2023). Equation (6) adds that the first-order term reaches only problems the recipient solves with probability strictly between zero and one. RLVR sharpens the recipient's pass probabilities toward zero and one along its own stage (Yue et al., 2025). The factor $\sqrt{p_x(1 - p_x)}$ in the bound therefore shrinks as K grows for every donor, and at full dose the second-order remainder, which can also remove problems the recipient already solves, determines the outcome.

The two propositions are an accounting identity and a first-order bound, and together they turn the three questions into testable predictions, one per question.

- ❶ Each update raises probability on the problems its teacher solves, so near the checkpoint that formed it, the update should reproduce the gains and losses of that teacher's students.
- ❷ The acquisition rate of a fixed donor update should fall along the recipient's stage, while forfeiture of recipient-only problems persists. The mature recipient then exchanges its problems for the teacher's.
- ❸ The per-token advantage of distillation is positive on sampled tokens the teacher prefers and negative on tokens only the student prefers, so distillation should keep forfeiting recipient-only problems as K grows and acquire teacher-only problems at a falling rate. The average retains the recipient-only problems through its half of the own update (Allen-Zhu & Li, 2023), so averaging should overtake distillation once $\mathcal{U}_R$ is large.

4 Merging and Distillation along the Recipient's Stage

4.1 Updates Carry Their Teacher's Identity

The teacher of an OPD job determines what its update does. Across the nine jobs of the student-by-teacher matrix, the teacher column explains 98% of the variance of the endpoints on $\mathcal{E}_{\text{in}}$ and 96% on $\mathcal{E}_{\text{ext}}$ (Table 5). On average, students distilled from the teacher of their own lineage end 0.13 points below the off-diagonal mean, so sharing a lineage with the teacher gives no advantage. Although the lineages differ only in their seed, students distilled from the best RLVR teacher end about 7 points above those distilled from the worst on $\mathcal{E}_{\text{in}}$, and every student follows the same ordering of teachers. The OPD update of a job, ported to the SFT checkpoint of the lineage that supplied neither its student nor its teacher, gains and loses problems that overlap more with those of that SFT checkpoint distilled under the same teacher than under a third teacher: 1.6 to 1.9 times as much for the gained problems (Jaccard 0.34–0.48 against 0.19–0.29) and consistently more for the lost ones (0.52–0.56 against 0.41–0.48; Table 6). Strength alone cannot explain the shared losses.

The update also carries the teacher's gain to checkpoints of other lineages. The RLVR update adds 5.9 points to the SFT checkpoint of another lineage on both evaluation sets, and it adds 3.6 points on $\mathcal{E}_{\text{in}}$ even to the base checkpoint, which skipped SFT (Table 7). The OPD update adds 7.0 points

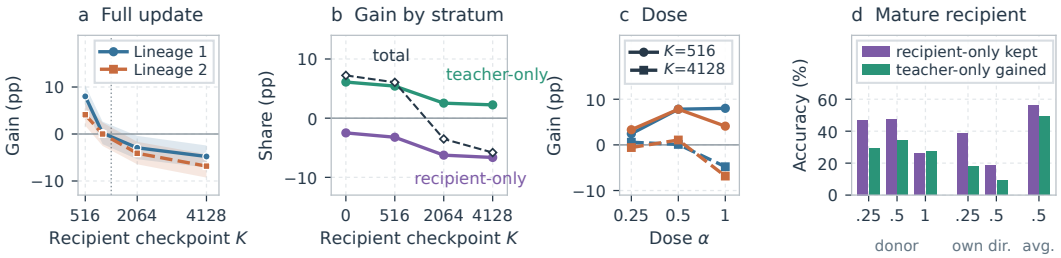


Figure 2: Early and mature recipients respond differently to the same updates on $\mathcal{E}_{\text{in}}$, with 95% bands in a; Figure 7 gives panels a and c on $\mathcal{E}_{\text{ext}}$. (a) The full-dose gain of Δ_s^{R} decreases at every checkpoint and changes sign at the dotted line. (b) In its weight-times-rate split, averaged over the two directions, the teacher-only share falls by more than half while the recipient-only loss more than doubles and pulls the total below zero. (c) The recipient at $K = 516$ gains at every dose, the one at $K = 4128$ at none. (d) On the mature recipient of lineage 1, averaged over both sets, the donor update exchanges strata. The own-direction extension keeps fewer recipient-only and gains fewer teacher-only problems than an equal-norm donor step, and the average (avg.) keeps more of each.

on $\mathcal{E}_{\text{in}}$ and 5.9 on $\mathcal{E}_{\text{ext}}$ to the SFT checkpoint of the third lineage. Their direction carries the gain. On the base and SFT checkpoints, the singular-value projection keeps the gain, and a sign shuffle brings it within half a point of zero. Negation turns it into a loss of 6 to 12 points. The OPD update reproduces the teacher's behavior without copying the teacher's weights. Its cosine with the teacher's RLVR update is 0.43, and a fit on that update and the difference between the two SFT checkpoints leaves 82% of its energy unexplained, averaged over the six directed pairs (Appendix I).

Takeaway 1: a post-training update carries its teacher's identity to other models.

On an early checkpoint of another lineage, an update changes many of the same problems its teacher changes in its own students, and the effect follows the update's direction, since a sign shuffle removes it.

4.2 The Same Update Helps Early and Hurts Late

The RLVR update of lineage s stays fixed and is added at full dose to the four saved checkpoints $R_t[K]$ of lineage t, and at a quarter and a half dose to the first and the mature one. Each gain is measured against the checkpoint the update is added to, and the controls act on the mature recipient.

Moving the recipient reverses the sign of the same update. The fixed RLVR update gains less at each successive checkpoint of another lineage, from a gain of 8.0 points on the first checkpoint of lineage 1 to a loss of 4.8 points on its mature checkpoint (Figure 2a), for both directed pairs and on both evaluation sets; at 1.7B and 8B the update also turns from a large gain on the SFT checkpoint into a loss on every mature recipient (Figure 4a, Table 1). The strata show why (Figure 2b; rates in Table 15): the update forfeits between 57% and 81% of the recipient-only problems at every stage, while it acquires 48% to 59% of the teacher-only problems on the first checkpoint and only 26% to 30% on the mature one. Meanwhile the teacher-only weight falls and the recipient-only weight rises until both lie between 6% and 9% (Table 15). On strata of similar size, a forfeiture rate of about 70% against an acquisition rate of about 30% gives a net loss.

The controls on size, own direction, and origin all place the source of the loss in the recipient. Smaller doses leave accuracy within noise and still trade recipient-only problems for teacher-only ones (Figure 2c,d). Extending the recipient's own update by the norm of half the donor costs 9.1 and 11.4 points on the two evaluation sets for lineage 1 and 4.1 and 3.1 for lineage 2, while half of the donor itself moves accuracy by about one point or less (Tables 8 to 10). For lineage 1 the own-direction extension keeps far fewer recipient-only problems than a donor step of the same norm, about 19% against 47% at half the donor's norm, and also gains fewer teacher-only ones (panel d). The OPD update formed on lineage s changes sign in the same way, from a gain of 8.5 points on the SFT checkpoint to a loss of 5.0 points on the mature one (Table 9), so the sign follows the recipient.

Halving the recipient's own update, the fusion of Liu et al. (2026), lets half of the donor add 8.3 and 6.2 points on $\mathcal{E}_{\text{in}}$ (Table 10); Section 5.1 tests this at scale.

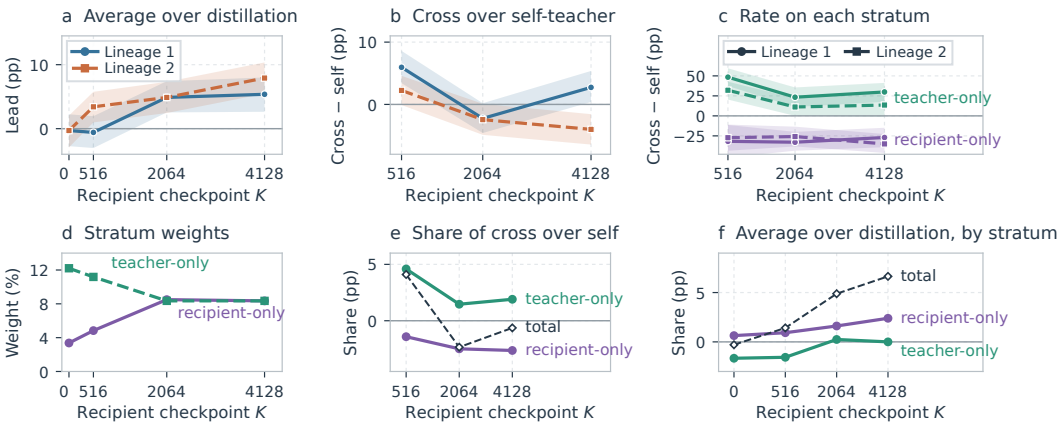


Figure 3: Averaging and distillation separate along the recipient's stage on $\mathcal{E}_{\text{in}}$, with 95% bands in a–c and means over the two directions in d–f. (a) Averaging and distillation are close early, and the average leads by 5 to 8 points from $K = 2064$ on. (b) The cross teacher's lead over the self-teacher falls from $K = 516$ to $K = 2064$ in both lineages. (c, d) The recipient-only loss stays near thirty points while the teacher-only gain falls to $K = 2064$ and partly recovers. The teacher-only weight falls and the recipient-only weight rises. (e, f) The teacher-only share of the cross teacher's lead shrinks toward the recipient-only loss. The average's lead over distillation grows on the recipient-only stratum, and its teacher-only deficit closes.

Takeaway 2: the same update helps an early recipient and hurts a mature one.

As the recipient's own reinforcement learning proceeds, a donor update gains fewer of the teacher's problems while still losing the recipient's own, so its benefit turns into a loss for RLVR and OPD updates.

4.3 Averaging Keeps What Distillation Forfeits

At each recipient checkpoint $K \in \{0, 516, 2064, 4128\}$ we compare the average A_K with one pass of distillation from $R_t[K]$ under the cross teacher R_s, each against its control, the shortened own update H_K and the self-teacher, which exists from $K = 516$ (Figure 3).

At 0.6B the two operations agree on most early recipients and separate where addition changes sign. From $K = 2064$ on, the average leads distillation by 3 to 8 points in every cell, while the cross teacher's lead over the self-teacher falls from up to 6.0 points at $K = 516$ to between -4.0 and $+2.7$ points on the mature recipients (Figure 3a,b). On the mature recipients at 1.7B and 8B, the average leads distillation in all eight cells, with the interval above zero in six (mean in Figure 4b).

The separation comes from the same exchange of problems that addition shows. Distillation forfeits a near-constant fraction of the recipient-only problems at every stage, and under the first protocol the six-pair mean from the mature checkpoints shows the same forfeiture (Table 11). Its acquisition of teacher-only problems falls from $K = 516$ to $K = 2064$ and only partly recovers on the mature recipient. By the stage split, the falling rate and the shifting weights both pull its total down (Figure 3c,d). On the mature recipients the cross teacher changes total accuracy by a few points relative to the self-teacher, yet it gains 14 to 34 points on the teacher-only problems and loses 27 to 35 on the recipient-only ones (Table 15). Averaging keeps the recipient-only problems through its half of the own update, and its lead over distillation in the recipient-only rate grows to 27–34 points at the mature checkpoint. Of the mature lead of 3.9 to 7.9 points, the recipient-only stratum contributes 1.9 to 2.8 and problems neither model solves most of the rest (Table 15).

Controls isolate the mechanism of each operation. The average exceeds its shortened-own control in all sixteen cells, so the donor's half contributes to its gain beyond the retreat toward SFT. Distillation, by contrast, moves against the recipient's own update: its displacement has a negative cosine with that update in all 28 layers. It approaches the teacher in behavior while less than a fifth of its displacement points toward the teacher's weights (Appendix I). The positive-only variant, which zeroes negative token advantages, still forfeits most of the recipient-only problems (Figure 5), so the forfeiture persists without the negative signal. Within the pass, the lead does not depend on when

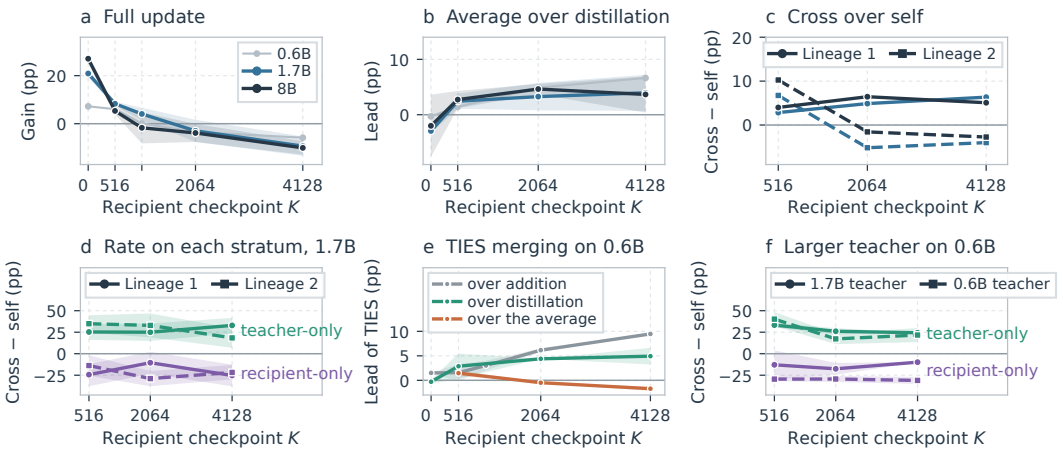


Figure 4: The results for addition and averaging recur at 1.7B and 8B and under TIES merging, and the distillation results recur in part. All panels are on $\mathcal{E}_{\text{in}}$. Panels a, b, and e show means over the two lineages with bands for their range, and panels a and b add the 0.6B pair in grey. Panels c and d show each lineage, and panel f each teacher. Color in panel c marks the model size as in panel a. (a) At both sizes the full update turns from a gain of 19 to 29 points on the SFT checkpoint to a loss on the mature one. (b) The mean lead of the average over distillation rises from below zero at $K = 0$ at both sizes. (c) The cross teacher's lead falls and turns negative in lineage 2, whose teacher is the weaker mature checkpoint, and stays positive in lineage 1. (d) At 1.7B the cross teacher still gains teacher-only and loses recipient-only problems in both lineages. (e) The lead of TIES over plain addition grows with the stage, and its lead over distillation is positive on the mature recipients. Its lead over the average falls below zero from $K = 2064$ on. (f) The 1.7B teacher forfeits fewer recipient-only problems than the 0.6B teacher, with strata defined against each teacher. Bands in d are 95% intervals, and panel f shows means over the two lineages with bands for their range.

distillation stops. From $K = 2064$ on, the average is ahead at every checkpoint saved during the pass, and the exchange of strata is present from the first save (Appendix F.1, Table 14).

Takeaway 3: averaging overtakes distillation as the recipient's own solved problems accumulate.

Distillation trades the recipient's own problems for the teacher's, whereas averaging keeps half of the recipient's own update and with it more of those problems.

5 The Own Update, the Teacher, and the Operator

5.1 The Recipient's Own Update

We test whether the recipient's own update limits what a donor adds on the mature 1.7B and 8B recipients, scaling the two updates separately, $S_t + a\,u_t + b\,\Delta$ with $a \in \{\frac{1}{2}, 1\}$ and $b \in \{0, \frac{1}{2}, 1\}$, for the RLVR donor Δ_s^{R} and the OPD donor $\Delta_{s,s}^{\text{O}}$. We measure the benefit of the donor, $B(a, b) = Y(S_t + a\,u_t + b\,\Delta) - Y(S_t + a\,u_t)$, and the interaction $I(b) = B(1, b) - B(\frac{1}{2}, b)$.

Lowering the dose mostly removes the loss, whereas halving the own update produces a gain (Table 1). With the full own update the full donor loses 3.5 to 13.4 points in all sixteen cells, and half of the donor gains with an interval excluding zero in three cells. With the own update halved, the same half donor gains up to 11.7 points in all sixteen cells, fourteen with intervals excluding zero. The interaction is negative in all thirty-two cells, thirty with intervals excluding zero, and reaches 15 points at the full dose. The more of its own update the recipient keeps, the less a donor adds. Averaging keeps half of each, the configuration with the largest B in Table 1.

5.2 Teacher Strength and Merging Operator

The accuracy of the teacher relative to the recipient predicts whether distillation helps. Under the first protocol, over six directed pairs, the net effect of distillation follows the teacher's margin over

Table 1: Keeping the full own update lowers the benefit of a donor on the mature 1.7B and 8B recipients, in points. Y is the recipient's accuracy, and the margin is the donor lineage's accuracy minus Y. $B(a, b)$ is listed for own scale a and donor scale b with gained / lost problems in grey. The interaction is $I(b) = B(1, b) - B(\frac{1}{2}, b)$, so $B(\frac{1}{2}, 1) = B(1, 1) - I(1)$. TIES columns compare with the full donor added (add) and with A_{4128} (avg). Shading marks intervals that exclude zero; Table 16 adds the 4B pair.

		Mature recipient		Benefit of the donor, own / donor scale			Interaction		TIES minus	
Size	Donor, direction	Y	margin	1 / 1	1 / $\frac{1}{2}$	$\frac{1}{2}$ / $\frac{1}{2}$	$I(\frac{1}{2})$	$I(1)$	add	avg
$\mathcal{E}_{\text{in}}$ (1,024 problems)										
1.7B	RLVR, 1←2	30.3	+7.2	$-5.3_{89/143}$	$+2.6_{116/89}$	$+11.7_{196/76}$	−9.1	−14.3	+12.9	−1.9
	RLVR, 2←1	37.5	−7.2	$-13.4_{63/200}$	$-0.3_{100/103}$	$+5.2_{139/86}$	−5.5	−10.6	+14.6	−1.9
	OPD, 1←2	30.3	+7.2	$-4.4_{88/133}$	$+3.8_{118/79}$	$+9.7_{170/71}$	−5.9	−12.2	+12.1	+0.3
	OPD, 2←1	37.5	−7.2	$-13.3_{61/197}$	$-0.7_{100/107}$	$+4.3_{129/85}$	−5.0	−10.6	+12.9	−2.5
8B	RLVR, 1←2	53.0	+5.0	$-7.2_{72/146}$	$+0.5_{104/99}$	$+1.8_{105/87}$	−1.3	−7.0	+14.6	+0.3
	RLVR, 2←1	58.0	−5.0	$-12.9_{58/190}$	$-0.4_{86/90}$	$+11.6_{172/53}$	−12.0	−15.1	+13.8	−2.4
	OPD, 1←2	53.0	+5.0	$-8.6_{71/159}$	$+0.3_{97/94}$	$+1.5_{108/93}$	−1.2	−7.7	+13.6	−1.9
	OPD, 2←1	58.0	−5.0	$-8.6_{69/157}$	$-0.6_{87/93}$	$+10.8_{162/51}$	−11.4	−10.8	+8.0	−3.1
$\mathcal{E}_{\text{ext}}$ (2,960 problems)										
1.7B	RLVR, 1←2	36.3	+3.1	$-8.1_{111/352}$	$-0.2_{181/187}$	$+7.1_{344/134}$	−7.3	−12.0	+10.9	−2.1
	RLVR, 2←1	39.4	−3.1	$-10.7_{120/438}$	$-1.7_{180/231}$	$+5.0_{303/154}$	−6.8	−9.7	+10.0	−2.6
	OPD, 1←2	36.3	+3.1	$-7.1_{126/335}$	$-1.3_{175/214}$	$+5.9_{328/152}$	−7.3	−11.2	+9.2	−1.7
	OPD, 2←1	39.4	−3.1	$-10.6_{112/426}$	$-1.6_{182/230}$	$+4.6_{306/170}$	−6.2	−9.8	+10.1	−2.0
8B	RLVR, 1←2	51.6	−1.2	$-5.7_{135/303}$	$+0.1_{181/179}$	$+5.9_{317/143}$	−5.8	−7.0	+9.1	−2.1
	RLVR, 2←1	50.3	+1.2	$-4.1_{157/277}$	$+2.0_{228/169}$	$+7.7_{355/126}$	−5.7	−6.6	+8.5	−1.8
	OPD, 1←2	51.6	−1.2	$-5.1_{139/291}$	$-0.9_{164/191}$	$+4.2_{288/164}$	−5.1	−6.6	+7.7	−1.3
	OPD, 2←1	50.3	+1.2	$-3.5_{161/266}$	$+2.6_{225/147}$	$+6.6_{325/131}$	−3.9	−6.2	+6.7	−1.9

the recipient, with a correlation of 0.98 on $\mathcal{E}_{\text{in}}$ and 0.99 on $\mathcal{E}_{\text{ext}}$ (Table 11). On $\mathcal{E}_{\text{in}}$ at 1.7B and 8B the cross teacher's lead falls and turns negative in the lineage with the weaker teacher and stays positive in the other (Figure 4c). Distilled into the 0.6B recipients, the 1.7B teacher keeps its lead over the self-teacher positive in every cell yet lets it fall from $K = 516$ to $K = 2064$ (Table 15; Figure 7h). It acquires teacher-only problems like the same-size teacher but forfeits far fewer recipient-only ones on average, about 10 to 18 points against about 30, with strata defined against each teacher (Figure 4f). Distillation from it therefore matches or beats averaging throughout (Table 15).

TIES scores 6.7 to 14.6 points above addition and 3.1 below to 0.3 above averaging on the mature 1.7B and 8B recipients (Table 1), and it beats distillation on every mature 0.6B recipient (Figure 4e).

6 Conclusion and Discussion

Post-training updates behave as portable task vectors with their teacher's identity, yet their value depends on how far the recipient has trained. Plasticity toward a donor falls along the recipient's RLVR stage, and the retained own update limits what a donor adds. Averaging overtakes distillation from a same-size peer as the recipient matures, and a stronger teacher lets distillation match averaging.

Merging acts on weights, distillation on behavior. Distillation pulls the recipient's answers toward the teacher's on every problem, including those only the recipient solves. Averaging keeps half of the recipient's own update, retains more of those problems, and solves some that neither checkpoint solves, in line with the view that an ensemble combines the features its members learn separately (Allen-Zhu & Li, 2023).

Choosing the operation and the peer. At 0.6B, once the teacher-only stratum shrinks to the size of the recipient-only one, addition loses and averaging leads same-size distillation. At 1.7B and 8B the average leads earlier. Distillation pays off when the teacher's margin is large, and the seed alone spreads the students of peer teachers by 7 points on $\mathcal{E}_{\text{in}}$.

AI USE STATEMENT

In this work, we used generative AI tools only to polish the writing of the manuscript. We have not used generative AI tools for developing the theoretical model or conceptual framework, formulating mathematical claims, designing the methodology or experiments, implementing the method, or interpreting the results, and the remaining tasks with required disclosure are not applicable to this work. We have reviewed all AI-assisted text and take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies mathematics post-training of publicly released language models on publicly available problem sets. It involves no human subjects, personal data, or sensitive content. The trained checkpoints are evaluated only on mathematics problems, and we identify no concern under the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Section 3 fixes the models, the two data sets, the objectives, the operations and their controls, the evaluation sets, and the strata. Appendix A lists the two training protocols, and Appendix I defines the transforms and the bootstrap protocol. The appendix lists the per-cell values behind the tables of the paper, including every dose, own-direction, porting, and average cell, every distillation arm from the mature 0.6B recipients, and the realized geometry of every structural transform; the values behind figure-only results (the 1.7B and 8B stage curves, the larger teacher's stratum rates, TIES along the stage, and the within-pass dynamics) are released with the code. The results concern Qwen3 models on mathematics prompts and two evaluation sets, with three lineages at 0.6B, one pair each at 1.7B and 8B, and a 4B pair as a boundary case. Appendix J states the scope. Intervals describe uncertainty over problems and exclude variation over training trajectories. The 0.6B stage results use one pair of lineages, and the six-pair contrasts extend the mature distillation contrast to three lineages (Table 11).

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Uuf2q9TfXGA.

Andrii Balashov. Reinforcement learning fine-tunes a sparse subnetwork in large language models. *arXiv preprint arXiv:2507.17107*, 2025. URL https://arxiv.org/abs/2507.17107.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *International Conference on Machine Learning*, volume 267, pp. 10818–10838, 2025. URL https://proceedings.mlr.press/v267/chu25c.html.

Yi Ding and Ruqi Zhang. Does OPD really distill? From noisy teacher to self-improvement. *arXiv preprint arXiv:2608.31046*, 2026. URL https://arxiv.org/abs/2608.31046.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632:768–774, 2024.

Shiyuan Feng, Huan-ang Gao, Haohan Chi, Hanlin Wu, Zhilong Zhang, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Weak-to-strong generalization via direct on-policy distillation. *arXiv preprint arXiv:2607.05394*, 2026. URL https://arxiv.org/abs/2607.05394.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, 2020.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Hugging Face. OpenR1-Math-220k. https://huggingface.co/datasets/open-r1/OpenR1-Math-220k, 2025.

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*, 2023.

Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. Model Stock: All we need is just a few fine-tuned models. In *European Conference on Computer Vision*, pp. 207–223, 2024. doi: 10.1007/978-3-031-72784-9_12. URL https://doi.org/10.1007/978-3-031-72784-9_12.

Feiyang Kang, Michael Kuchnik, Karthik Padthe, Marin Vlastelica, Ruoxi Jia, Carole-Jean Wu, and Newsha Ardalani. Quagmires in SFT-RL post-training: When high SFT scores mislead and what to use instead. In *International Conference on Learning Representations*, volume 2026, pp. 56876–56918, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5d6ae8ba43ecb378030753c4408ef9bd-Paper-Conference.pdf.

Boyan Li, Bingsen Chen, Chenghao Yang, Ping Nie, Chen Zhao, and Xi Ye. Sequential beats joint: On the interplay between on-policy distillation and RLVR. *arXiv preprint arXiv:2609.04108*, 2026a. URL https://arxiv.org/abs/2609.04108.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large

language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026b. URL https://arxiv.org/abs/2604.13016.

Zhaoyi Li, Deyang Kong, Yuan Wei, Evan Yang, Ranran Shen, Mahardika Krisna Ihsani, Ming Yang, Wei Zhang, Chuan Hao, Jian Yang, Ran Tao, Bryan Dai, Shikun Zhang, Wei Ye, Ying Wei, and Defu Lian. Every coin has two sides: On the dual nature of generalization in on-policy distillation of large language models. *arXiv preprint arXiv:2608.16647*, 2026c. URL https://arxiv.org/abs/2608.16647.

Runze Liu, Jiashun Liu, Xu Wan, Yuqian Fu, and Ling Pan. When RL fails after SFT: Rejuvenating model plasticity for robust SFT-to-RL handoff. *arXiv preprint arXiv:2606.09932*, 2026. URL https://arxiv.org/abs/2606.09932.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tur, and Hao Peng. Reinforcement learning finetunes small subnetworks in large language models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 132119–132138, 2025. doi: 10.52202/085713-4399. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/bf235a1d6780afd979f2f81676f43413-Paper-Conference.pdf.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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/2402.03300.

Zhennan Shen, Yanshu Li, Qingyu Yin, Chak Tou Leong, Zhilin Wang, Yanxu Chen, Rongduo Han, Sunbowen Lee, and Yi R. Fung. On the geometry of on-policy distillation. *arXiv preprint arXiv:2606.07082*, 2026. URL https://arxiv.org/abs/2606.07082.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. In *International Conference on Learning Representations*, 2026.

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*, pp. 1279–1297, 2025. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026. URL https://arxiv.org/abs/2604.00626.

Yuanyi Wang, Yanggan Gu, Yiming Zhang, Qi Zhou, Zhaoyi Yan, Congkai Xie, Xinyao Wang, Jianbo Yuan, and Hongxia Yang. Model merging scaling laws in large language models. In *International Conference on Machine Learning*, 2026a.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026b. URL https://arxiv.org/abs/2605.26844.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model Soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, 2022.

Chenrui Wu, Zexi Li, Jiajun Bu, Jiangchuan Liu, and Haishuai Wang. Sparsity curse: Understanding RLVR model parameter space from model merging. *arXiv preprint arXiv:2606.18521*, 2026. URL https://arxiv.org/abs/2606.18521.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Fangxu Yu, Weijia Xu, Michael Xu, Tianyi Zhou, and Zinan Lin. Weak-to-strong on-policy distillation. *arXiv preprint arXiv:2607.26246*, 2026a. URL https://arxiv.org/abs/2607.26246.

Guo Yu, Wenlin Liu, Yulan Hu, Hao-Xuan Ma, Jun-Peng Jiang, and Han-Jia Ye. Dense supervision, sparse updates: On the sparsity and geometry of on-policy distillation. *arXiv preprint arXiv:2606.13657*, 2026b. URL https://arxiv.org/abs/2606.13657.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 113222–113244, 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a4277440d50f1f15d2cb4c14f7e0c0d2-Paper-Conference.pdf.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 57654–57689, 2025. doi: 10.52202/085713-1933. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/537d5aa768c2d534016a4d06f87bc8fb-Paper-Conference.pdf.

Luca Zhou, Daniele Solombrino, Donato Crisostomi, Maria Sofia Bucarelli, Giuseppe Alessio D'Inverno, Fabrizio Silvestri, and Emanuele Rodolà. On task vectors and gradients. In *Proceedings of UniReps: the Third Edition of the Workshop on Unifying Representations in Neural Models*, volume 322, pp. 398–417, 2026. URL https://proceedings.mlr.press/v322/zhou26a.html.

A Protocols and Implementation Note

Table 2: The two training protocols. Both use the same 14,448 prompts in the same archived order, eight rollouts per prompt, a fresh optimizer per job, learning rate 10^{-6}, reasoning mode off, and 2,048-token responses.

	first protocol	second protocol
prompt batch	7	28
passes over the prompts	2	1
optimizer updates	4,128	2,064 (516 collections)
checkpoints	every 516 updates	collections 129, 258, 387, 516
OPD loss	original implementation	corrected normalization
used by	matrix, $R_r[K]$, six-pair contrasts	distillation jobs of Section 4.3

Table 2 compares the two protocols. SFT, shared by both, runs 516 optimizer updates at batch size 56 and learning rate 10^{-6}. Under the second protocol a collection is one prompt batch of 28 prompts, which the trainer consumes in four optimizer updates. The first protocol uses an implementation of the sampled-token distillation loss in which the distillation term of the first microbatch on each of the seven actor ranks reads the global token normalization of the preceding optimizer update instead of the current one, so 7 of the 56 responses in every update are normalized with stale statistics. RLVR is unaffected because its loss refreshes the normalization before aggregation. The second protocol refreshes the normalization on every microbatch. The calibration of one OPD job of the first protocol under the corrected loss raises the mean over saved checkpoints by 0.16 points on $\mathcal{E}_{\text{in}}$ and lowers it by 0.01 points on $\mathcal{E}_{\text{ext}}$, with terminal differences of mixed sign and mostly below half a point. Every OPD comparison in the paper is made within one implementation. The recipient checkpoints $R_r[K]$, the student-by-teacher matrix, and the six-pair contrasts of Section 4.3 come from the first protocol. The distillation jobs of Section 4.3, including the positive-only and random-deletion arms, come from the second. The positive-only arm zeroes negative token advantages, and the random-deletion arm drops tokens at random as a control for losing signal.

B Notation and Operations

Table 3 lists the operations with their formulas and sections, and Table 4 collects the symbols. Contrasts reported with intervals follow the convention of Table 15. Entries whose 95% problem-level interval excludes zero are shaded, light blue for positive and light yellow for negative values with a depth that follows the magnitude, and entries whose interval contains zero are set in grey.

Table 3: Unified view of the operations on updates studied in the paper. Every operation acts on archived checkpoints of a directed pair (t, s) of lineages; τ denotes the structural transforms: identity, projection onto the singular directions carrying 90% of the squared singular values, a random sign flip of every coordinate, and negation.

Operation	Formula	Recipient	Section
addition, dose α	$R_t[K] + \alpha\,\Delta_s^{\mathrm{R}}$, $\alpha \in \{0.25, 0.5, 1\}$	S_t, θ_0, $R_t[K]$	4.1, 4.2
structural transform	$\theta + \tau(\Delta_s^{\mathrm{R}})$, $\theta + \tau(\Delta_{r,s}^{\mathrm{O}})$	θ_0, S_q	4.1
own-direction extension	$R_t + \alpha\,(\lVert\Delta_s^{\mathrm{R}}\rVert / \lVert\Delta_t^{\mathrm{R}}\rVert)\,\Delta_t^{\mathrm{R}}$	R_t	4.2
ported OPD update	$\theta + \Delta_{s,s}^{\mathrm{O}}$	S_t, R_t	4.1, 4.2
shortened own update	$H_K = S_t + \frac{1}{2}u_t(K)$	$R_t[K]$	4.2, 4.3
average	$A_K = S_t + \frac{1}{2}u_t(K) + \frac{1}{2}\Delta$	$R_t[K]$	4.2, 4.3
TIES merging	trimmed, sign-aligned merge of $u_t(K)$ and Δ on S_t	$R_t[K]$	5.2
composition	$S_t + a\,u_t + b\,\Delta$, $a \in \{\frac{1}{2}, 1\}$, $b \in \{0, \frac{1}{2}, 1\}$	R_t (1.7B, 8B)	5.1
distillation	$R_t[K] \to \mathrm{OPD}(R_s)$, one pass, second protocol	$R_t[K]$	4.3

Table 4: Notation of the main text; lineage indices r, s, t refer to the lineages of one pair or of the three 0.6B lineages, and q denotes the 0.6B lineage outside a pair.

Symbol	Definition
Checkpoints	
θ_0	base checkpoint
S_r	SFT checkpoint of lineage r
$R_r[K]$, R_r	RLVR checkpoint of lineage r after K updates, with $R_r[0] = S_r$ and the mature $R_r = R_r[4128]$
$O_{r\to s}$	OPD endpoint of a student started from S_r under the frozen teacher R_s
O_K, O_K^{self}	one-pass distillation endpoints from $R_t[K]$ under the cross teacher R_s and under the self-teacher $R_t[K]$
collection	one prompt batch of the second protocol, four optimizer updates
Updates	
$u_t(K) = R_t[K] - S_t$	own update of the recipient at stage K
$\Delta_s^{\text{R}} = R_s - S_s$, $\Delta_{r,s}^{\text{O}} = O_{r\to s} - S_r$	RLVR update of lineage s, and OPD update of student r under teacher s
$\Delta, T = R_s$	donor update and its teacher
$h_{r,s} = S_s - S_r$	difference between two SFT checkpoints
$\tau(\Delta)$	identity, projection, shuffle, or negation of an update (Appendix I)
Operations	
$\theta + \alpha\Delta$	addition at dose α
$A_K = S_t + \frac{1}{2}u_t(K) + \frac{1}{2}\Delta$	average
$H_K = S_t + \frac{1}{2}u_t(K)$	shortened own update, the control for averaging
$S_t + a\,u_t + b\,\Delta$	composition with own scale a and donor scale b (Section 5.1)
cross, self	distillation from $R_t[K]$ under R_s, and under the frozen $R_t[K]$
Evaluation and estimands	
$\mathcal{E}_{\text{in}}$, $\mathcal{E}_{\text{ext}}$	evaluation sets of 1,024 and 2,960 problems
$Y(\theta)$, g	greedy accuracy in percentage points, and the change of accuracy of an operation
$\mathcal{U}_T, \mathcal{U}_R, \mathcal{U}_B, \mathcal{U}_N$	teacher-only, recipient-only, both-solved, neither-solved
w_σ, r_σ	stratum weight and rate (Definition 2)
G, E, D, C	stage contrasts: addition gain, cross minus self-teacher, average minus cross teacher, average minus H_K
$B(a, b)$, $I(b)$	benefit of the donor and its interaction with the own scale
$\hat{a}_j(x, y)$	per-token OPD advantage, Equation (2)

C PROOFS

Proof of Proposition 1. The four strata partition $\mathcal{X}$, so $Y(\theta') - Y(\theta) = \frac{100}{N}\sum_\sigma \sum_{x\in\mathcal{U}_\sigma}\big(c_{\theta'}(x) - c_\theta(x)\big) = \sum_\sigma \frac{|\mathcal{U}_\sigma|}{N} r_\sigma$, which is Equation (4). For two stages, $g' - g = \sum_\sigma (w'_\sigma r'_\sigma - w_\sigma r_\sigma)$; adding and subtracting $\sum_\sigma w'_\sigma r_\sigma$ and $\sum_\sigma w_\sigma r'_\sigma$ inside the sum and collecting terms gives Equation (5). □

Proof of Proposition 2. The expansion is Taylor's theorem with the Lagrange remainder applied to $\alpha \mapsto J(\theta + \alpha\Delta)$, which is twice differentiable because π_θ is. For the bound, the likelihood-ratio identity gives $\nabla p_x = \mathrm{E}_y[v\,\nabla \log \pi_\theta(y \mid x)]$, and since $\mathrm{E}_y[\nabla \log \pi_\theta] = 0$ the same expectation equals $\mathrm{E}_y[(v - p_x)\nabla \log \pi_\theta]$. Hence $\langle \nabla p_x, \Delta\rangle = \mathrm{E}_y[(v - p_x)\,\langle \nabla \log \pi_\theta, \Delta\rangle]$, and the Cauchy–Schwarz inequality bounds it by $\sqrt{\mathrm{E}_y[(v - p_x)^2]}\sqrt{\mathrm{E}_y[\langle \nabla \log \pi_\theta, \Delta\rangle^2]}$. For a binary reward $\mathrm{E}_y[(v - p_x)^2] = p_x(1 - p_x)$, and the second factor is $\Delta^\top F_x \Delta$. □

D PER-CELL VALUES FOR THE TEACHER'S IDENTITY

D.1 TEACHER IDENTITY

Table 5 lists the terminal accuracy of the nine OPD jobs of the student-by-teacher matrix, and Table 6 the overlap of the problems that a ported OPD update gains and loses with those of the endpoints trained under the same and under a third teacher. The two tables carry the evidence of Section 4.1.

Table 5: The terminal greedy accuracy of the student-by-teacher OPD matrix follows the teacher column in every row. Cells give $\mathcal{E}_{\text{in}}$ and $\mathcal{E}_{\text{ext}}$ accuracy in percentage points, and diagonal cells pair each student with the teacher of its own lineage.

Student	Teacher R_1	Teacher R_2	Teacher R_3
S_1	14.0 / 18.6	17.7 / 20.4	9.9 / 16.2
S_2	13.7 / 19.1	17.2 / 20.0	10.9 / 16.7
S_3	14.6 / 18.3	17.1 / 20.9	10.4 / 16.9
teacher-column share	98.40% / 95.84% of variance		
diagonal minus off-diagonal mean	-0.13 / -0.11 points		

Table 6: The ported OPD update $\Delta^{\text{O}}_{s,s}$ on S_t shares more of its gained and lost problems with the endpoint trained under the same teacher, $O_{t\to s}$, than with the endpoint trained under a third teacher, $O_{t\to q}$. Agreement is the share of problems on which the ported update and the endpoint have the same greedy correctness; shading marks the same-teacher column.

		agreement (%)		gained Jaccard		lost Jaccard	
Direction	Set	same	third	same	third	same	third
$\Delta^{\text{O}}_{2,2} \to S_1$	$\mathcal{E}_{\text{in}}$	87.30	84.96	0.405	0.217	0.524	0.481
$\Delta^{\text{O}}_{2,2} \to S_1$	$\mathcal{E}_{\text{ext}}$	89.12	85.61	0.430	0.264	0.522	0.408
$\Delta^{\text{O}}_{1,1} \to S_2$	$\mathcal{E}_{\text{in}}$	87.01	84.08	0.340	0.190	0.519	0.475
$\Delta^{\text{O}}_{1,1} \to S_2$	$\mathcal{E}_{\text{ext}}$	90.47	85.81	0.480	0.286	0.561	0.425

D.2 PORTABILITY AND DOSE

Table 7 gives the addition gain of the RLVR and OPD updates on the SFT and base checkpoints of the other lineages under the identity, projection, shuffle, and negation transforms, and Table 8 the gain at each dose on the first and the mature checkpoint.

Table 7: Post-training updates keep their gain when added to checkpoints that did not produce them. The singular-value projection keeps the gain, and a sign shuffle removes it. Entries are the addition gain $g(\theta;\tau(\Delta))$ of Equation (3) averaged over the six directed pairs of three lineages (three donors for the base-checkpoint rows), in percentage points with 95% problem-level intervals.

Update Δ	Target θ	Transform τ	$\mathcal{E}_{\text{in}}$		$\mathcal{E}_{\text{ext}}$	
Δ^{R}_{s}	S_t	identity	$+5.88$	[4.54, 7.29]	$+5.90$	[5.04, 6.72]
Δ^{R}_{s}	θ_0	identity	$+3.61$	[1.73, 5.57]	$+2.02$	[0.91, 3.15]
Δ^{R}_{s}	θ_0	projection (90% per matrix)	$+4.66$	[2.60, 6.71]	$+1.75$	[0.57, 2.85]
Δ^{R}_{s}	θ_0	sign shuffle	-0.39	[-1.92, 1.11]	-0.26	[-1.15, 0.59]
Δ^{R}_{s}	θ_0	negation	-8.11	[-10.03, -6.28]	-11.51	[-12.71, -10.25]
$\Delta^{\text{O}}_{r,s}$	S_q	identity	$+7.00$	[5.60, 8.43]	$+5.91$	[5.10, 6.79]
$\Delta^{\text{O}}_{r,s}$	S_q	projection (90% per matrix)	$+6.46$	[5.11, 7.88]	$+6.12$	[5.31, 6.99]
$\Delta^{\text{O}}_{r,s}$	S_q	sign shuffle	0.00	[-0.75, 0.80]	$+0.08$	[-0.37, 0.54]
$\Delta^{\text{O}}_{r,s}$	S_q	negation	-6.07	[-7.29, -4.87]	-9.70	[-10.65, -8.72]

Table 8: The donor update gains at every dose on the early recipient and at no dose on the mature recipient. Entries are problems gained / lost / net against the same target without the update, with the net in percentage points and its 95% problem-level interval; $\mathcal{E}_{\text{in}}$ has 1,024 and $\mathcal{E}_{\text{ext}}$ 2,960 problems.

Direction	Target	α	$\mathcal{E}_{\text{in}}$		$\mathcal{E}_{\text{ext}}$	
$\Delta_2^{\text{R}} \to$ lineage 1	$R_1[516]$	0.25	71/46/+25, +2.44	[0.39, 4.49]	187/128/+59, +1.99	[0.81, 3.18]
$\Delta_2^{\text{R}} \to$ lineage 1	$R_1[516]$	0.5	121/41/+80, +7.81	[5.37, 10.16]	320/142/+178, +6.01	[4.63, 7.43]
$\Delta_2^{\text{R}} \to$ lineage 1	R_1	0.25	68/62/+6, +0.59	[−1.66, 2.73]	161/150/+11, +0.37	[−0.81, 1.52]
$\Delta_2^{\text{R}} \to$ lineage 1	R_1	0.5	73/72/+1, +0.10	[−2.25, 2.44]	175/150/+25, +0.84	[−0.37, 2.06]
$\Delta_1^{\text{R}} \to$ lineage 2	$R_2[516]$	0.25	81/47/+34, +3.32	[1.17, 5.47]	203/142/+61, +2.06	[0.81, 3.28]
$\Delta_1^{\text{R}} \to$ lineage 2	$R_2[516]$	0.5	126/46/+80, +7.81	[5.37, 10.35]	253/143/+110, +3.72	[2.36, 5.03]
$\Delta_1^{\text{R}} \to$ lineage 2	R_2	0.25	72/78/−6, −0.59	[−2.93, 1.76]	171/168/+3, +0.10	[−1.15, 1.32]
$\Delta_1^{\text{R}} \to$ lineage 2	R_2	0.5	87/76/+11, +1.07	[−1.27, 3.52]	178/179/−1, −0.03	[−1.28, 1.18]

E Per-Cell Values for Early and Mature Recipients

Table 9 lists the mature-recipient interventions of Section 4.2, the donor update at three doses, the own-direction extension, and the ported OPD update, for lineage 1. Table 10 completes both recipients with the average and the shortened own update. Table 11, in Appendix F, covers the six directed pairs of the three lineages.

Table 9: The same update loses on the mature recipient at full dose and not at smaller doses. The own-direction extension loses at every dose, and the OPD update flips sign between the SFT and the mature checkpoint. Entries are addition gains in percentage points against the same target, with 95% problem-level intervals where paired on one target.

Update	Target	$\mathcal{E}_{\text{in}}$		$\mathcal{E}_{\text{ext}}$	
$0.25\,\Delta_2^{\text{R}}$	R_1	+0.59	[−1.66, 2.73]	+0.37	[−0.81, 1.52]
$0.5\,\Delta_2^{\text{R}}$	R_1	+0.10	[−2.25, 2.44]	+0.84	[−0.37, 2.06]
Δ_2^{R}	R_1	−4.79	[−7.13, −2.44]	−4.80	[−6.08, −3.51]
own dir., norm of $0.25\,\Delta_2^{\text{R}}$	R_1	−3.81	[−5.96, −1.66]	−2.87	[−3.99, −1.76]
own dir., norm of $0.5\,\Delta_2^{\text{R}}$	R_1	−9.08	[−11.33, −6.74]	−11.42	[−12.77, −10.03]
$\Delta_{2,2}^{\text{O}}$	S_1	+8.50	[6.25, 10.74]	+7.13	[5.78, 8.48]
$\Delta_{2,2}^{\text{O}}$	R_1	−4.98	[−7.32, −2.64]	−4.56	[−5.84, −3.24]

Table 10: The own-direction extension, the ported OPD update, the average, and the shortened own update H_{4128} on both recipients, complementing the lineage-1 cells of Table 9. Entries are problems gained / lost / net or the net in percentage points against the stated target, with 95% problem-level intervals where the comparison is paired on one target. The last block compares the average with H_{4128} and so isolates the contribution of the donor's half.

Update	Target	$\mathcal{E}_{\text{in}}$		$\mathcal{E}_{\text{ext}}$	
own direction, norm of $0.25\,\Delta_1^{\text{R}}$	R_2	67/83/−16, −1.56	[−3.91, 0.88]	148/189/−41, −1.39	[−2.60, −0.17]
own direction, norm of $0.5\,\Delta_1^{\text{R}}$	R_2	57/99/−42, −4.10	[−6.45, −1.56]	126/217/−91, −3.07	[−4.29, −1.86]
$\Delta_{1,1}^{\text{O}}$	S_2	114/43/+71, +6.93	[4.59, 9.28]	314/122/+192, +6.49	[5.14, 7.87]
$\Delta_{1,1}^{\text{O}}$	R_2	53/109/−56, −5.47	[−7.91, −3.03]	109/295/−186, −6.28	[−7.60, −5.00]
average $S_2 + \frac{1}{2}\Delta_2^{\text{R}} + \frac{1}{2}\Delta_1^{\text{R}}$	R_2	108/68/+40		234/147/+87	
average $S_1 + \frac{1}{2}\Delta_1^{\text{R}} + \frac{1}{2}\Delta_2^{\text{R}}$	R_1	125/55/+70		267/113/+154	
$H_{4128} = S_1 + \frac{1}{2}\Delta_1^{\text{R}}$	R_1	−1.46	[−3.91, 0.98]	+1.08	[−0.20, 2.37]
$H_{4128} = S_2 + \frac{1}{2}\Delta_2^{\text{R}}$	R_2	−2.25	[−4.69, 0.29]	−0.74	[−2.03, 0.54]
average of lineage 1	H_{4128}	+8.30	[5.66, 10.84]	+4.12	[2.77, 5.47]
average of lineage 2	H_{4128}	+6.15	[3.71, 8.59]	+3.68	[2.33, 5.03]

F Per-Cell Values for Averaging and Distillation

Table 12 lists every cell of the design of Section 4.3: the correct problems of the recipient, its shortened own update, the average, and the two distillation endpoints at each checkpoint, with the

Table 11: Under the first protocol, the cross-teacher OPD trajectories from the three mature recipients reallocate correctness toward the teacher-only problems, averaged over the six directed pairs. Entries are cross-minus-self contrasts averaged over the saved checkpoints of the two-pass trajectory, in percentage points with 95% problem-level intervals, and the terminal gain of the cross-teacher endpoint over its start.

Quantity	$\mathcal{E}_{\text{in}}$		$\mathcal{E}_{\text{ext}}$	
cross minus self, recipient-only	−23.06	[−26.88, −19.29]	−27.41	[−29.95, −24.87]
cross minus self, teacher-only	+23.85	[20.63, 27.11]	+26.68	[24.20, 29.10]
cross teacher, terminal gain over the start	−0.78		−0.01	
correlation of terminal gain with teacher margin	0.98		0.99	

three contrasts. Table 13 compares the cross teacher with the self-teacher on the strata, Table 14 lists every distillation arm of the two mature recipients, and Figure 5 draws the four arms over the pass.

Table 12: Every cell of the averaging-or-distillation design. Columns give, for each evaluation set, lineage t, and recipient checkpoint K, the correct problems of the recipient $R_t[K]$, the shortened own update H_K, the average A_K, and the cross-teacher and self-teacher distillation endpoints after one pass, followed by the contrasts $D = Y(A_K) - Y(O_K)$, $E = Y(O_K) - Y(O_K^{\text{self}})$, and $C = Y(A_K) - Y(H_K)$ in percentage points with 95% problem-level intervals. No self-teacher arm exists at $K = 0$. $\mathcal{E}_{\text{in}}$ has 1,024 problems and $\mathcal{E}_{\text{ext}}$ 2,960.

Set	t	K	$R_t[K]$	H_K	A_K	cross	self	D		E		C	
$\mathcal{E}_{\text{in}}$	1	0	72	72	177	180	–	−0.29	[−2.73, 2.25]	–		+10.25	[8.01, 12.70]
$\mathcal{E}_{\text{in}}$	1	516	94	93	165	171	110	−0.59	[−3.03, 1.95]	+5.96	[3.42, 8.59]	+7.03	[4.69, 9.38]
$\mathcal{E}_{\text{in}}$	1	2064	171	135	205	155	178	+4.88	[2.54, 7.42]	−2.25	[−4.59, 0.20]	+6.84	[4.30, 9.47]
$\mathcal{E}_{\text{in}}$	1	4128	153	138	223	168	140	+5.37	[2.64, 8.01]	+2.73	[0.39, 5.37]	+8.30	[5.66, 10.94]
$\mathcal{E}_{\text{in}}$	2	0	76	76	150	153	–	−0.29	[−2.93, 2.05]	–		+7.23	[4.88, 9.47]
$\mathcal{E}_{\text{in}}$	2	516	105	92	160	125	102	+3.42	[0.98, 5.76]	+2.25	[−0.10, 4.59]	+6.64	[4.20, 9.08]
$\mathcal{E}_{\text{in}}$	2	2064	161	146	191	141	166	+4.88	[2.44, 7.32]	−2.44	[−4.88, 0.10]	+4.39	[2.05, 7.03]
$\mathcal{E}_{\text{in}}$	2	4128	176	153	216	135	176	+7.91	[5.47, 10.35]	−4.00	[−6.45, −1.56]	+6.15	[3.71, 8.79]
$\mathcal{E}_{\text{ext}}$	1	0	382	382	596	601	–	−0.17	[−1.42, 1.15]	–		+7.23	[5.91, 8.58]
$\mathcal{E}_{\text{ext}}$	1	516	427	382	606	604	427	+0.07	[−1.18, 1.42]	+5.98	[4.66, 7.36]	+7.57	[6.18, 8.89]
$\mathcal{E}_{\text{ext}}$	1	2064	638	551	697	608	633	+3.01	[1.59, 4.36]	−0.84	[−2.16, 0.54]	+4.93	[3.55, 6.25]
$\mathcal{E}_{\text{ext}}$	1	4128	567	599	721	607	573	+3.85	[2.47, 5.20]	+1.15	[−0.20, 2.47]	+4.12	[2.84, 5.47]
$\mathcal{E}_{\text{ext}}$	2	0	370	370	586	524	–	+2.09	[0.81, 3.38]	–		+7.30	[6.01, 8.58]
$\mathcal{E}_{\text{ext}}$	2	516	478	416	598	564	468	+1.15	[−0.24, 2.47]	+3.24	[1.86, 4.63]	+6.15	[4.83, 7.50]
$\mathcal{E}_{\text{ext}}$	2	2064	623	537	703	532	634	+5.78	[4.49, 7.09]	−3.45	[−4.76, −2.09]	+5.61	[4.22, 7.06]
$\mathcal{E}_{\text{ext}}$	2	4128	622	600	709	560	592	+5.03	[3.75, 6.35]	−1.08	[−2.40, 0.34]	+3.68	[2.36, 5.00]

Table 13: The cross teacher lowers recipient-only accuracy and raises teacher-only accuracy relative to the frozen self-teacher at every checkpoint in both directions, while the overall contrast follows the sign of the teacher margin. Entries are point estimates of cross minus self in percentage points on $\mathcal{E}_{\text{in}}$ / $\mathcal{E}_{\text{ext}}$ at each saved collection of the pass. The strata of direction $1 \leftarrow 2$ contain $74/188$ recipient-only and $97/243$ teacher-only problems, and those of $2 \leftarrow 1$ the reverse.

Direction	Quantity	col. 129	258	387	516
$1 \leftarrow 2$	overall	+1.37 / +0.61	+3.13 / +1.86	+4.10 / +1.96	+2.73 / +1.15
$1 \leftarrow 2$	recipient-only	−14.86 / −31.91	−12.16 / −21.28	−4.05 / −25.00	−27.03 / −26.60
$1 \leftarrow 2$	teacher-only	+23.71 / +31.28	+26.80 / +32.51	+34.02 / +32.51	+29.90 / +23.46
$2 \leftarrow 1$	overall	−5.47 / −1.69	−2.93 / −1.72	−3.22 / −1.89	−4.00 / −1.08
$2 \leftarrow 1$	recipient-only	−44.33 / −26.75	−35.05 / −29.63	−31.96 / −30.45	−35.05 / −31.28
$2 \leftarrow 1$	teacher-only	+4.05 / +26.60	+18.92 / +25.00	+24.32 / +27.13	+13.51 / +33.51

Table 14: Every distillation arm from the two mature recipients at the four saved collections of the second protocol, as correct problems on $\mathcal{E}_{\text{in}}$ / $\mathcal{E}_{\text{ext}}$. The recipient-only stratum of direction $1 \leftarrow 2$ contains $74/188$ problems and the teacher-only stratum $97/243$. Direction $2 \leftarrow 1$ has the reverse. The start of every arm is the mature checkpoint itself, at $153/567$ for R_1 and $176/622$ for R_2.

Student	Arm	Stratum	col. 129	258	387	516
R_1	cross (R_2)	overall	161/599	168/622	173/623	168/607
R_1	self (R_1)	overall	147/581	136/567	131/565	140/573
R_1	positive-only	overall	180/633	181/623	160/592	163/600
R_1	random deletion	overall	178/599	186/627	187/605	180/632
R_1	cross (R_2)	recipient-only	23/62	26/76	23/66	17/62
R_1	self (R_1)	recipient-only	34/122	35/116	26/113	37/112
R_1	positive-only	recipient-only	19/74	21/84	19/67	22/66
R_1	random deletion	recipient-only	22/69	21/71	25/69	20/75
R_1	cross (R_2)	teacher-only	44/132	41/135	50/135	46/129
R_1	self (R_1)	teacher-only	21/56	15/56	17/56	17/72
R_1	positive-only	teacher-only	42/135	48/124	49/124	42/124
R_1	random deletion	teacher-only	44/117	50/132	50/133	52/139
R_2	cross (R_1)	overall	124/568	145/551	138/553	135/560
R_2	self (R_2)	overall	180/618	175/602	171/609	176/592
R_2	positive-only	overall	144/546	145/576	148/574	141/561
R_2	random deletion	overall	141/559	133/562	143/557	132/556
R_2	cross (R_1)	recipient-only	13/75	15/66	18/62	21/58
R_2	self (R_2)	recipient-only	56/140	49/138	49/136	55/134
R_2	positive-only	recipient-only	29/67	21/69	29/71	21/71
R_2	random deletion	recipient-only	24/75	20/69	22/70	19/69
R_2	cross (R_1)	teacher-only	25/105	32/105	33/108	30/116
R_2	self (R_2)	teacher-only	22/55	18/58	15/57	20/53
R_2	positive-only	teacher-only	30/91	35/105	29/105	34/102
R_2	random deletion	teacher-only	34/95	28/103	40/103	28/109

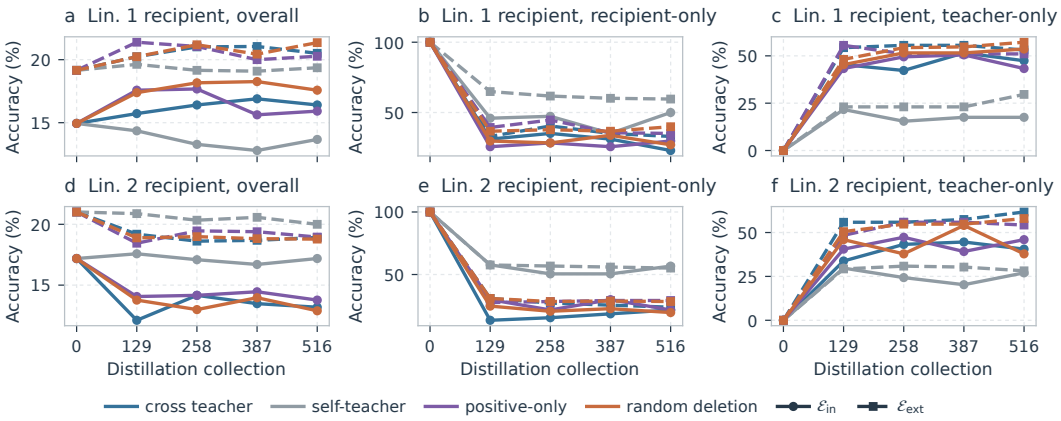


Figure 5: Over one pass from each mature recipient, the cross teacher lowers the recipient-only stratum and raises the teacher-only stratum relative to the self-teacher at every checkpoint in both directions, and the positive-only arm loses most of the recipient-only stratum.

F.1 Stage Contrasts and Within-Pass Dynamics

Table 15 lists the contrasts of the three operations at $K \in \{516, 2064, 4128\}$ with their stratum rates and weights. Figure 6 follows the distillation jobs from $K \in \{0, 516, 2064\}$ through the pass, and Tables 13 and 14 and Figure 5 do so from $K = 4128$. From $K = 2064$ the average leads the cross teacher at every save, and the cross teacher gains teacher-only problems and forfeits recipient-only ones from the first save on, that is, within the first 516 optimizer updates of the pass. The values themselves still move between saves, and the cross teacher's lead over the self-teacher changes sign within the pass for lineage 1 from $K = 2064$.

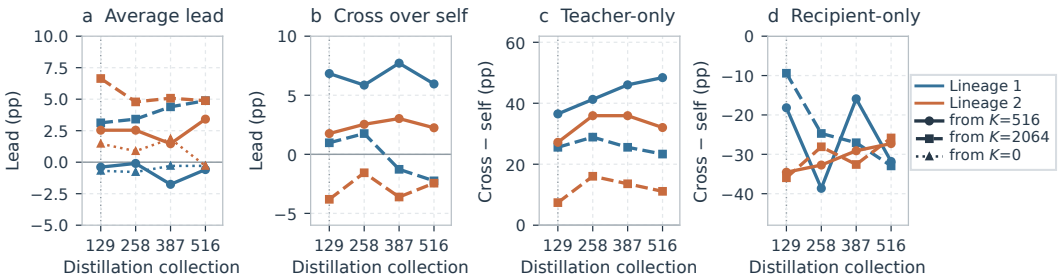


Figure 6: Contrasts at every checkpoint saved during the distillation pass on $\mathcal{E}_{\text{in}}$, for passes from three recipient checkpoints. One collection is one prompt batch of four optimizer updates, and a checkpoint is saved every 129 collections. (a) From $K = 2064$ the average leads the cross teacher at every save. (b) The cross teacher's lead over the self-teacher moves within the pass and changes sign for lineage 1 from $K = 2064$. (c, d) The cross teacher's teacher-only gain and recipient-only loss are present from the first save on.

Table 15: The stage contrasts change along one axis, the recipient checkpoint K, for both directions of the pair (recipient ← teacher) and both evaluation sets, in percentage points. Stratum cells give the accuracy change on the stratum with its weight in percent in small grey type, so weight times rate is its share of the total (Equation (4)). Shading marks 95% problem-level intervals that exclude zero, light blue for positive and light yellow for negative values, deeper for larger magnitudes; grey entries have intervals that contain zero. The last row is the difference of two point estimates, D minus the larger-teacher row, and carries no interval.

	$K = 516$				$K = 2064$				$K = 4128$			
	$1 \leftarrow 2$		$2 \leftarrow 1$		$1 \leftarrow 2$		$2 \leftarrow 1$		$1 \leftarrow 2$		$2 \leftarrow 1$	
Operation and stratum	$\mathcal{E}_{\text{in}}$	$\mathcal{E}_{\text{ext}}$	$\mathcal{E}_{\text{in}}$	$\mathcal{E}_{\text{ext}}$	$\mathcal{E}_{\text{in}}$	$\mathcal{E}_{\text{ext}}$	$\mathcal{E}_{\text{in}}$	$\mathcal{E}_{\text{ext}}$	$\mathcal{E}_{\text{in}}$	$\mathcal{E}_{\text{ext}}$	$\mathcal{E}_{\text{in}}$	$\mathcal{E}_{\text{ext}}$
Addition (G): $R_t[K] + \Delta_s^{\text{R}}$ minus $R_t[K]$												
total	$+8.0$	$+5.1$	$+4.1$	$+1.7$	-2.9	-3.9	-4.1	-4.7	-4.8	-4.8	-6.8	-6.5
teacher-only	$+48.4_{12}$	$+58.7_{11}$	$+48.5_{10}$	$+53.2_{8}$	$+30.0_{9}$	$+35.5_{7}$	$+30.9_{8}$	$+34.5_{6}$	$+25.8_{9}$	$+28.8_{8}$	$+28.4_{7}$	$+30.3_{6}$
recipient-only	-56.8_{4}	-80.6_{5}	-74.5_{5}	-72.3_{5}	-72.9_{8}	-67.0_{8}	-74.2_{9}	-69.6_{8}	-79.7_{7}	-68.1_{6}	-79.4_{9}	-73.3_{8}
Distillation (E): cross teacher minus self-teacher, both started from $R_t[K]$												
total	$+6.0$	$+6.0$	$+2.2$	$+3.2$	-2.2	-0.8	-2.4	-3.4	$+2.7$	$+1.1$	-4.0	-1.1
teacher-only	$+48.4_{12}$	$+44.4_{11}$	$+32.0_{10}$	$+36.3_{8}$	$+23.3_{9}$	$+19.0_{7}$	$+11.1_{8}$	$+19.5_{6}$	$+29.9_{9}$	$+23.5_{8}$	$+13.5_{7}$	$+33.5_{6}$
recipient-only	-31.8_{4}	-33.6_{5}	-27.3_{5}	-22.3_{5}	-32.9_{8}	-28.6_{8}	-25.8_{9}	-37.0_{8}	-27.0_{7}	-26.6_{6}	-35.1_{9}	-31.3_{8}
Average against distillation (D): A_K minus cross teacher												
total	-0.6	$+0.1$	$+3.4$	$+1.1$	$+4.9$	$+3.0$	$+4.9$	$+5.8$	$+5.4$	$+3.9$	$+7.9$	$+5.0$
teacher-only	-19.8_{12}	-10.0_{11}	-6.8_{10}	-7.2_{8}	$+0.0_{9}$	$+2.8_{7}$	$+6.2_{8}$	$+1.1_{6}$	-4.1_{9}	$+2.1_{8}$	$+5.4_{7}$	-7.4_{6}
recipient-only	$+6.8_{4}$	$+18.7_{5}$	$+29.1_{5}$	$+18.2_{5}$	$+17.6_{8}$	$+23.8_{8}$	$+20.2_{9}$	$+22.6_{8}$	$+27.0_{7}$	$+29.3_{6}$	$+29.9_{9}$	$+34.2_{8}$
neither-solved	$+1.7_{79}$	$+0.7_{74}$	$+2.0_{80}$	$+1.4_{76}$	$+5.0_{75}$	$+1.6_{71}$	$+2.0_{76}$	$+4.0_{73}$	$+4.8_{76}$	$+1.6_{73}$	$+4.8_{76}$	$+3.2_{73}$
both-solved	$+4.0_{5}$	-1.7_{10}	$+20.0_{5}$	-2.4_{11}	-3.5_{8}	-1.2_{14}	$+15.3_{7}$	$+7.9_{13}$	$+2.5_{8}$	$+5.3_{13}$	$+13.9_{8}$	$+3.2_{13}$
Average against its shortened-own control (C): A_K minus H_K												
total	$+7.0$	$+7.6$	$+6.6$	$+6.1$	$+6.8$	$+4.9$	$+4.4$	$+5.6$	$+8.3$	$+4.1$	$+6.2$	$+3.7$
Larger teacher: the 1.7B cross teacher minus the 0.6B cross teacher, both started from $R_t[K]$												
total	$+3.7$	$+5.4$	$+13.0$	$+7.5$	$+6.6$	$+5.1$	$+10.2$	$+9.6$	$+4.9$	$+3.8$	$+11.8$	$+10.4$
A_K minus 1.7B teacher	-4.3	-5.4	-9.6	-6.4	-1.8	-2.1	-5.3	-3.9	$+0.5$	$+0.0$	-3.9	-5.4

G Curves on the External Set

Figure 7 repeats the main $\mathcal{E}_{\text{in}}$ curves on $\mathcal{E}_{\text{ext}}$. The sign change of the full update and the lead of the average from $K = 2064$ on recur on the external problems at 0.6B and 1.7B. The cross teacher's lead falls from $K = 516$ to $K = 2064$ at 0.6B and in lineage 2 at 1.7B.

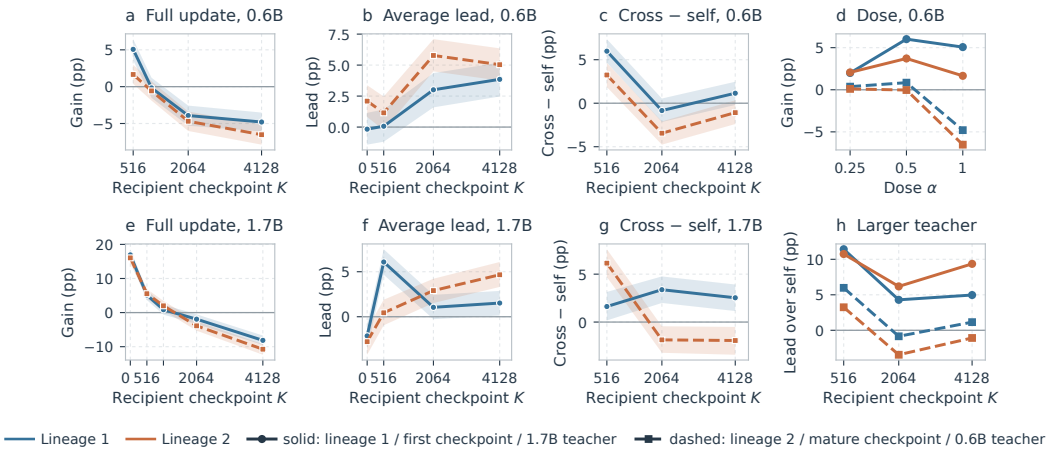


Figure 7: Panels a and c of Figure 2, panels a and b of Figure 3, the 1.7B curves of Figure 4, and the larger teacher's lead over the self-teacher, on the external set $\mathcal{E}_{\text{ext}}$, with 95% problem-level intervals in a–c and e–g. (a–d) The 0.6B pair: the full update changes sign, the average leads distillation from $K = 2064$ on, the cross teacher's lead falls from $K = 516$ to $K = 2064$, and the mature recipient gains at no dose. (e–g) The 1.7B pair, where the cross teacher's lead falls in lineage 2 and not in lineage 1. (h) The larger teacher on the 0.6B recipients.

H Other Model Sizes and Implementation Details

Table 16: The composition matrix of Table 1 for the 4B pair, whose lineage-1 RLVR checkpoint collapses at $K = 4128$ (37 of 1,024 problems on $\mathcal{E}_{\text{in}}$). Each direction therefore pairs a donor far weaker or far stronger than its recipient. Entries are in percentage points on both evaluation sets.

	$\mathcal{E}_{\text{in}}$							$\mathcal{E}_{\text{ext}}$						
Donor and direction	$B(1,1)$	$B(1,\frac{1}{2})$	$B(\frac{1}{2},\frac{1}{2})$	$I(\frac{1}{2})$	$I(1)$	TIES−add	TIES−avg	$B(1,1)$	$B(1,\frac{1}{2})$	$B(\frac{1}{2},\frac{1}{2})$	$I(\frac{1}{2})$	$I(1)$	TIES−add	TIES−avg
Qwen3-4B														
RLVR donor, $1 \leftarrow 2$	+0.3	+0.5	+7.1	−6.6	−2.5	+15.3	−14.8	−2.2	−1.7	+4.7	−6.4	−2.6	+24.5	−7.6
RLVR donor, $2 \leftarrow 1$	−49.2	−21.5	−23.3	+1.9	+3.0	+13.9	−15.5	−40.9	−13.7	−9.4	−4.4	−1.2	+23.9	−7.4
OPD donor, $1 \leftarrow 2$	−0.3	+0.2	+5.2	−5.0	−3.0	+14.0	−14.8	−1.8	−1.5	+4.6	−6.0	−2.1	+21.4	−10.0
OPD donor, $2 \leftarrow 1$	−48.6	−15.3	−24.0	+8.7	+3.2	+15.7	−12.4	−36.2	−8.2	−11.2	+2.9	+3.6	+20.7	−4.1

The 4B pair is a boundary case of the design. Its lineage-1 RLVR trajectory reaches 481 of 1,024 problems on $\mathcal{E}_{\text{in}}$ at $K = 2064$ and 37 at $K = 4128$, while lineage 2 reaches 536 at $K = 4128$. At the mature stage each lineage therefore meets a donor that is either far weaker or far stronger than itself. The full update of the strong lineage neither gains nor loses on the collapsed recipient (a change of 0.3 points on $\mathcal{E}_{\text{in}}$ whose interval contains zero), while the update of the collapsed lineage lowers the strong recipient by 49 points. In the 4B rows of Table 16 the interaction is negative in seven of sixteen cells, positive in three, and its interval contains zero in six. Here a positive interaction marks a smaller loss instead of a gain. The OPD donor $2 \leftarrow 1$ accounts for all three positive cells, and on $\mathcal{E}_{\text{in}}$ its half donor lowers accuracy by 15 points on the full own update and by 24 points on the half own update. The counts of the main text exclude this pair.

The 1.7B and 8B pairs use the data and protocols of Appendix A with the base checkpoint replaced and follow the 0.6B design cell by cell. Their distillation jobs follow the second protocol. TIES merging is the mergekit implementation with per-tensor magnitude trimming at density 0.2, sign election by summed magnitude, disjoint mean, and unit scaling, applied to $u_t(K)$ and the donor update Δ on S_t; density 0.5 serves as a fixed sensitivity check and gives the same signs. The larger-teacher block distills the mature 1.7B checkpoint of the opposite lineage into the 0.6B recipients at $K \in \{0, 516, 2064, 4128\}$ under the second protocol, with the archived self-teacher jobs as controls.

I INTERVENTIONS AND GEOMETRY

Trajectory of distillation. Under the first protocol, each mature checkpoint is distilled for two passes under the mature checkpoint of another lineage over the six directed pairs, and every saved checkpoint is placed in the basis of the two RLVR updates and the difference between the two SFT checkpoints, with coordinates and reference points defined in the paragraph on trajectory coordinates below. The trajectory moves away from the sum and stays closer to the teacher than its start. The normalized distance to the sum exceeds one at all 48 readings. From the first saved checkpoint to the last, the coordinate on the recipient's own update falls by a fifth while the coordinate on the teacher's update doubles. The energy outside the span of the three basis directions grows from a third to three fifths of the displacement (Table 19). Table 18 gives the geometry of the one-pass jobs of the second protocol. Distillation therefore retracts part of the recipient's own update and approaches the teacher in function without interpolating between the two checkpoints in weight space.

Transforms. For a target checkpoint θ and an update Δ, an intervention materializes $\theta + \tau(\Delta)$ in the training precision and reports the addition gain $g(\theta; \tau(\Delta))$ of Equation (3) under greedy decoding. The identity transform gives the full update. The projection transform keeps, in every weight matrix, the leading singular directions whose squared singular values reach 90% of the total. The shuffle transform multiplies every coordinate of Δ by an independent random sign, which keeps the norm and the magnitude of every coordinate and destroys the sign structure. The negation transform returns $-\Delta$. Table 17 compares the realized update with the intended one. Dose, own-direction, average, and ported constructions are formed in single precision from the archived checkpoints and stored in the training precision, and their realized norms and cosines with the intended update are recorded with each model.

Table 17: The realized geometry of each structural transform after materialization. The projection keeps the direction, and the shuffle keeps the norm while removing the direction. The negation reverses the direction. Ranges cover all materialized models of both update families. The 90% threshold applies per weight matrix, and the realized projection keeps 98% of the total energy. Storage in the training precision separates its cosine from its norm ratio.

Transform	cosine with full update	norm ratio	squared-norm energy
projection	0.9726 to 0.9747	0.9897 to 0.9905	0.9796 to 0.9811
shuffle	0.0007 to 0.0013	0.9993 to 0.9994	0.9986 to 0.9988
negation	-0.9987	0.9987 to 0.9990	0.9974 to 0.9980

The two updates. For each pair (r, s), the fit $\Delta^{\mathrm{O}}_{r,s} \approx a\, h_{r,s} + b\, \Delta^{\mathrm{R}}_{s}$ gives $a = 0.030$ and $b = 0.411$ with residual energy 0.816, norm ratio 0.960, and cosine 0.429 between the two updates, each within a narrow range over the six pairs; the OPD update is neither a copy of the teacher's RLVR update nor an interpolation between that update and the SFT difference.

Geometry of the realized updates. Table 18 summarizes the parameter geometry of the realized updates of both lineages under the first protocol and of the mature distillation jobs under the second, computed from the stored checkpoints in single precision with tied weights counted once. Concentration is measured by the effective support $P_{\mathrm{eff}}/N = (\sum \delta^2)^2/(N \sum \delta^4)$ of an update δ, so smaller values mean energy concentrated on fewer coordinates. The SFT update is more concentrated and places 28% of its energy in the embeddings, whereas the mature RLVR update and the cross-teacher distillation displacement are more dispersed, of similar root-mean-square size, and spread across the transformer layers. The self-teacher displacement is half as large and favors the query, key, and MLP gate projections. The concentration of the recipient's own RLVR update falls along its stage, from P_{eff}/N of 6.4% at $K = 516$ to 15.0% at $K = 4128$. The cross-teacher displacement $d = O - R_t$ has negative cosine with the recipient's own update in all 28 layers and positive cosine with the donor update and with the teacher offset $\gamma_T = R_s - R_t$. Its projection coefficient onto γ_T grows from 0.15–0.17 to 0.28–0.29 over the pass. The line toward the teacher still carries only 16% to 18% of its energy, and the Euclidean distance to the teacher shrinks by 6% and 4%. Distillation therefore moves toward the teacher without interpolating along the line between the two checkpoints. The two mature RLVR updates of the pair have cosine 0.03 although 41% of their largest-magnitude

coordinates overlap, so they point in near-orthogonal directions on shared coordinates. For the static constructions, the increment $A_K - H_K$ computed in single precision keeps the sign of the nominal $\frac{1}{2}\Delta_s^{\mathrm{R}}$ on every coordinate with projection coefficient 1.0003 to 1.0005. Rounding to the training precision adds a residual, so the stored difference $A_K - H_K$ of the evaluated models has norm ratio 1.27 to 1.30 and cosine 0.77 to 0.79 with the nominal $\frac{1}{2}\Delta_s^{\mathrm{R}}$.

Table 18: Parameter geometry of the realized updates for lineages 1 and 2 (lineage 1 / lineage 2). Effective support is P_{eff}/N in percent, larger for a less concentrated update. RMS is the root-mean-square coordinate change in units of 10^{-5}. Embedding share is the fraction of update energy in the embeddings. Cosines are taken with the recipient's own mature update $u_t = R_t - S_t$ and with the teacher offset $\gamma_T = R_s - R_t$ at the end of the pass.

Update	effective support	RMS	embedding share	$\cos(\cdot, u_t)$	$\cos(\cdot, \gamma_T)$
SFT $S_t - \theta_0$	10.6 / 10.7	5.55 / 5.55	0.278 / 0.277		
RLVR $R_t - S_t$	15.0 / 15.2	5.77 / 5.86	0.038 / 0.041	1	
RLVR $R_t[516] - S_t$	6.4 / 6.5				
cross-teacher distillation $O_{4128} - R_t$	14.9 / 14.9	5.72 / 5.74	0.026 / 0.028	-0.29 / -0.27	$+0.43$ / $+0.41$
self-teacher distillation $O^{\mathrm{self}}_{4128} - R_t$	9.0 / 8.3	2.87 / 2.64	0.052 / 0.058	$+0.003$ / $+0.003$	
cross distillation from $R_t[2064]$	14.0 / 14.1			-0.18 / -0.16	$+0.42$ / $+0.40$

Coordinates and distances of the trained trajectory. For a directed pair (t, s) and a checkpoint θ_k of the trajectory that starts from R_t, let $y_k = \theta_k - S_t$. The least-squares coefficients (a_k, b_k, c_k) of $y_k \approx a_k\Delta_t^{\mathrm{R}} + b_k\Delta_s^{\mathrm{R}} + c_k h_{t,s}$ place R_t at $(1, 0, 0)$, the sum at $(1, 1, 0)$, the ported point at $(0, 1, 0)$, and the teacher at $(0, 1, 1)$. The residual energy is $\|y_k - a_k\Delta_t^{\mathrm{R}} - b_k\Delta_s^{\mathrm{R}} - c_k h_{t,s}\|^2/\|y_k\|^2$, and the condition number of the basis stays below seven in every pair. The normalized distance to a reference point Q is $D_Q(k) = \|\theta_k - Q\|/\|Q - R_t\|$, equal to one at the start. Averaged over the six pairs, the coefficients are $(0.853, 0.170, 0.007)$ at the first checkpoint and $(0.674, 0.371, 0.028)$ at the last.

Table 19: Along the cross-teacher trajectories from the mature recipients under the first protocol, the normalized distance to the sum rises while the distances to the ported point and the teacher stay below one. The energy outside the three-vector span grows. Entries are equal-cell means over the six directed pairs.

checkpoint	D_{sum}	D_{ported}	D_{teacher}	residual energy
0	1.000	1.000	1.000	0.000
516	1.065	0.961	0.963	0.352
1032	1.099	0.949	0.952	0.452
1548	1.126	0.947	0.949	0.506
2064	1.151	0.948	0.950	0.542
2580	1.170	0.951	0.953	0.566
3096	1.189	0.955	0.957	0.587
3612	1.208	0.961	0.962	0.605
4128	1.227	0.967	0.967	0.620

Realized signal of the positive-only and random-deletion arms. The positive-only arm drops every token with a negative advantage, and the random-deletion arm drops tokens at random. Over the first 129 collections of recipient R_1 the positive-only arm keeps a median of 0.774 of the tokens with nonzero advantage in each microbatch and the random-deletion arm 0.396, with logged gradient norms 0.448 and 0.818. The two arms therefore differ in the realized signal as well as in which tokens they drop.

Bootstrap protocol. The bootstrap resamples evaluation problems within each set with replacement, using 10,000 draws for the dose interventions and 2,000 for the stage and distillation cells. All cells and contrasts of one comparison share the same draws, and equal weights over the lineages are applied after resampling where a quantity averages over lineages. For the six-pair trajectories the resampling is stratified by the four correctness strata of the archived endpoints so that recipient-only and teacher-only accuracies are resampled within their own item sets.

J SCOPE

The lineages are peers: they share base model, data, and protocol and differ only in their seed, which isolates the recipient's stage but leaves open how updates from different data or tasks compose. All training and evaluation use mathematics problems with greedy decoding, and all models belong to the Qwen3 family. The account of Section 3.4 rests on the recipient's pass probability p_x in Equation (6), which RLVR sharpens toward zero and one (Yue et al., 2025). Within the family, the stage regularities recur at 0.6B, 1.7B, and 8B (the 4B pair is a boundary case, Appendix H), for donors from reinforcement learning and from distillation, under addition, averaging, TIES merging, and distillation, and in part under a larger teacher; teacher identity and the controls of Section 3.2 are measured at 0.6B. The 0.6B stage results rest on one pair of lineages in both directions, and the intervals cover problems but not training trajectories. Stage and teacher margin also move together in each pair; the larger teacher and the two directions of each pair separate them only in part. The account of Section 3.4 makes a testable prediction beyond this scope. Reinforcement learning sharpens the recipient's pass probabilities on its own task. Updates from another domain should therefore meet problems on which the recipient is still uncertain, and their benefit should fade less along the recipient's stage than a same-task update.