DOES TRAINING ON A TEXT MAKE IT MORE LIKELY? NOT UNDER DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Likelihood-based data audits assume that a model assigns higher probability to text it was trained on. We test this assumption with a randomized-inclusion design, changing the prediction target while keeping the input prefixes fixed, and find that it fails under distillation. We randomly choose which problems enter training, score every trajectory before and after training, and compare the gain of the included trajectories with the gain of the excluded ones. Fine-tuning on the original tokens gives the included trajectories the larger gain. Using the teacher's argmax labels instead, however, reverses this advantage: the contrast in gains changes from $+0.016$ to -0.019 nats per token, and the reversal replicates in a second model family. It arises at the roughly 5% of positions where the teacher's label differs from the original token: training only at these positions reproduces the negative contrast, whereas training only at the remaining positions yields a positive one. To first order, the expected contrast is non-negative when training on the scored tokens and can take either sign when the target changes. For auditing, this reversal flips a common membership score, which ranks texts by how much their likelihood rose: the score separates included from excluded trajectories after fine-tuning and ranks them the wrong way round after distillation. A control in which the student is distilled from a copy of itself reverses the score just as much despite a far smaller contrast, so a reversed score alone cannot identify distillation. A randomized contrast in either direction shows that data was used; a null does not show that it was not.

1 INTRODUCTION

Many ways of asking what a language model was trained on rest on the same premise: training on a text makes that text more likely. Membership inference attacks score a candidate by its loss, or by its loss relative to a reference model (Carlini et al., 2022; Duan et al., 2024). Pretraining-data and contamination detectors look for tokens the model finds unusually easy (Shi et al., 2024; Zhang et al., 2025b). Memorization studies measure how much more likely a training example is than a held-out one (Carlini et al., 2023; Zhang et al., 2023). The sign of the effect is rarely questioned, because for the objective these tools were built around, maximum likelihood on the text itself, it is the direction the objective pushes.

Post-training has moved away from that objective. In distillation a student is trained on prefixes it generated itself, but the target at each position comes from a teacher (Agarwal et al., 2024; Gu et al., 2024; Li et al., 2026). Where the teacher agrees with the trajectory the update resembles fine-tuning; where it disagrees, the student is pushed away from the token that an auditor would later score. Work on distillation and privacy has asked whether the teacher's training data leaks into the student (Jagielski et al., 2023; Zhang et al., 2025c; Borkar et al., 2026). Whether the student's own training trajectories end up more likely or less likely has not been measured under a design that separates the effect of including a trajectory from the drift that training imposes on every trajectory.

We measure it with randomized inclusion. From one pool of self-generated reasoning trajectories we draw a random subset of problems into training, train on those only, and compare the change in trajectory likelihood between included and excluded problems (§2). Fine-tuning on the trajectories' own tokens behaves as the premise predicts: included trajectories gain $+0.016$ nats/token more than excluded ones. Training the same prefixes on a teacher's argmax labels reverses the sign, to -0.019,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

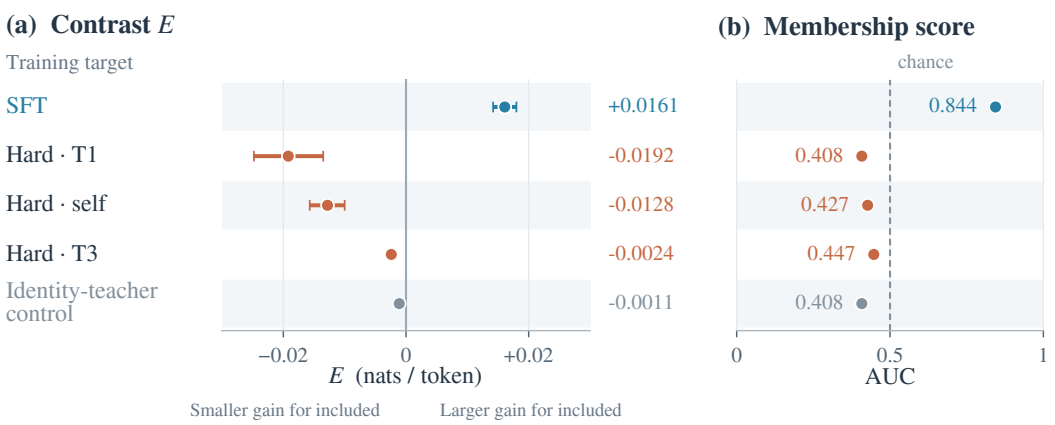


Figure 1: Training on a trajectory can lower its likelihood relative to trajectories that were not trained on. (a) The contrast E for fine-tuning (SFT), for hard distillation with teachers T1, self, and T3, and for the identity-teacher control; Qwen3-4B student, three assignments, t-intervals. (b) AUC of the reference-based membership score with the direction "became more likely $\Rightarrow$ member" fixed in advance (point estimates). The score detects fine-tuning and falls below chance for every other arm, including the control.

on the same problems and the same assignments (Figure 1). In this setting the included trajectories end up less likely than the trajectories excluded from this training stage; more generally, replacing the target shrinks the gain that inclusion confers.

Two accounts make different predictions here. If exposure alone drove the trace, the sign would not depend on the objective, and a negative reading would have to be noise or drift. If the trace follows what the loss asks the model to predict, the seen text gains likelihood only where it coincides with the target, and the sign depends on the objective. The data support the second account. The same exposure produces opposite signs under the two objectives. Replacing the target lowers the contrast in most settings and reverses it in some, including when the student's own argmax is the target and no external teacher is involved. The negative sign is carried by the positions where target and trajectory differ.

We support this from two sides. On the side of what carries the trace, the supervised labels gain likelihood in the same runs in which the trajectories lose it ($+0.013$ to $+0.018$; §4). On the side of where the negative sign comes from, it lives in the roughly 5% of positions at which the teacher's label differs from the trajectory's token. Supervising only those positions yields -0.029; supervising only the agreeing positions yields $+0.005$. A one-step lemma states why the leading term of the expected contrast is non-negative under fine-tuning on the scored text and why replacing the target removes that guarantee.

Under on-policy distillation, where every trajectory is resampled by the current student, preliminary runs show the same direction at smaller size: a fine-tuning positive control reads positive, the soft target reads negative, and the hard-target readings are negative in two seeds (§3). The consequence for audits is concrete. A reference-based membership score, which ranks candidates by how much their log-probability rose, detects fine-tuning (AUC 0.84) and ranks members below non-members after distillation (AUC 0.41–0.45). An identity-teacher control, in which the student is distilled from a copy of itself, reverses the score just as strongly (AUC 0.41) while its contrast E is a seventeenth of the T1 arm's. A reversed score alone therefore cannot identify distillation; the contrast E separates the two arms where the score does not (§5). For data that an auditor randomized into or out of training before release, a significant contrast in either direction shows that the data was used, and a null does not show non-use.

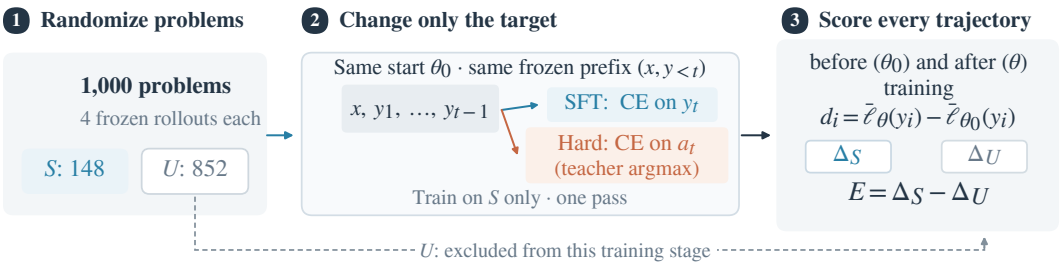


Figure 2: The randomized-inclusion design. A random subset of problems enters training; on the same frozen prefixes the loss targets either the trajectory's own token y_t (SFT) or the teacher's argmax a_t (hard distillation); every trajectory, included or not, is scored before and after training, and $E = \Delta_S - \Delta_U$. S and U denote the included and the excluded problems; Δ_S and Δ_U are the mean changes in trajectory log-probability within each.

Contributions.

- **Replacing the training target weakens, and can reverse, the trace of inclusion (§3).** Under randomized inclusion, fine-tuning raises the relative likelihood of the included trajectories and distilling the same prefixes onto a teacher's labels lowers it. Target replacement lowers the contrast in seven of the eight student settings we ran and reverses its sign for three SFT students from two model families. The reversal appears when the student's own argmax is the target, and preliminary on-policy experiments also yield negative contrasts.
- **The negative sign is produced at disagreement positions (§4).** Trajectory tokens and teacher labels move in opposite directions on the same prefixes, and restricting supervision to the 5% of positions where they differ reproduces the negative sign. Concurrent work locates the task benefit of on-policy distillation at a few positions (Wang et al., 2026b; Liu et al., 2026; Ding & Zhang, 2026); we measure what supervising those positions does to the trace that inclusion leaves on the whole trajectory.
- **Default-direction membership scores reverse after distillation, and the reversal is not evidence of distillation (§5).** Reference-based scores fall below chance after distillation and after the identity-teacher control alike; endpoint-only scores read nothing; scoring the teacher's labels recovers the signal. We state what a randomized likelihood audit can establish and what it cannot: use, never non-use.

2 Measuring what training does to its own data

Auditing setting. We take the position of a data owner who plants a randomized canary: a random subset of a set of candidate texts is released to a trainer and the rest is withheld, and the owner later queries the released model, and possibly a reference model, for per-token log-probabilities of every text. This is the setting of dataset inference and one-run privacy auditing (Steinke et al., 2023; Maini et al., 2024; Zhang et al., 2025a), and it is the setting in which a likelihood signal has a causal reading. The trainer's objective is unknown to the owner.

Randomized inclusion. We start from a pool of N=1000 math problems with four trajectories each, sampled from the initial student θ_0. For each of three assignments we draw m=148 problems (592 trajectories) into the training set S and leave the remaining 852 problems as U; all four trajectories of a problem move together. The student is trained on S for one pass, and every trajectory in $S \cup U$ is then scored by θ_0 and by the trained model θ with the teacher-forced prefixes that training used (Figure 2).

The contrast E. For trajectory i let $d_i = \bar{\ell}_\theta(y_i) - \bar{\ell}_{\theta_0}(y_i)$ be the change in its mean per-token log-probability, and let Δ_S and Δ_U be the means of d_i over included and excluded trajectories. We report the contrast $E = \Delta_S - \Delta_U$, the amount by which inclusion changes a trajectory's likelihood gain relative to exclusion. Because assignment is random, whatever training does to the model

Table 1: E (nats/token) by training target. Qwen3-4B-SFT student, self-generated pool, mean over three assignments with a t-interval (df $= 2$). The soft-target arms were trained with int8 optimizer state in an earlier round; all other rows use fp32 optimizer state.

Target	Teacher	E [95% CI]
SFT (own tokens)	–	$+0.0161$ $[+0.0142, +0.0180]$
Hard (argmax labels)	T1 (RL-tuned from θ_0)	-0.0192 $[-0.0248, -0.0135]$
Hard (argmax labels)	self (θ_0)	-0.0128 $[-0.0157, -0.0100]$
Hard (argmax labels)	T3 (8B, SFT recipe)	-0.0024 $[-0.0029, -0.0018]$
Soft (top-20 forward KL)	T1	-0.0038 $[-0.0064, -0.0013]$
Soft (top-20 reverse KL)	T1	-0.0037 $[-0.0063, -0.0010]$
Identity-teacher control	self	-0.0011 $[-0.0017, -0.0004]$
Paired difference, hard (T1) − SFT		-0.0353 $[-0.0427, -0.0278]$

that is not specific to the included problems moves Δ_S and Δ_U alike and cancels (Zhang et al., 2023; Steinke et al., 2023). The cancellation matters in practice: for the hard-distillation arm the drift Δ_U is -0.082 against an E of -0.019, so the raw likelihood change on training data mostly measures drift. E is an average contrast under the realized assignments, not a per-example effect, and a negative E means that included trajectories gained less than excluded ones, whether or not both gained.

Training targets. With the prefix fixed to the trajectory, only the target of the loss changes. **SFT** applies cross-entropy to the trajectory's own token y_t. **Hard distillation** applies cross-entropy to the teacher's full-vocabulary argmax a_t at the same prefix. **Soft distillation** minimizes a forward or reverse KL to the teacher on the student's top-20 support. The **identity-teacher control** runs soft distillation with a bf16 copy of the student as teacher; the two distributions differ only by numerical precision (the copy's argmax differs from the student's at 0.24% of positions), so it is a training run whose target does not depend on which problems were included. Three teachers supply the hard target: θ_0 after 100 steps of RL with verifiable rewards (T1), θ_0 itself (self), and an 8B model fine-tuned with the student's recipe (T3). All of these keep the scored trajectories as fixed prefixes, the regime of distillation on pre-sampled student rollouts (Agarwal et al., 2024).

The online estimand. When the student resamples its trajectories at every step, the pool's trajectories are never trained on and E is undefined. We then draw two rollouts per prompt at each step, train on one chosen by a coin, and after training score both with the final model, subtracting each rollout's log-probability at generation. The mean paired difference over 592 pairs from 363 prompt clusters is the *trained-versus-sibling contrast*; below we call it the online contrast. It is a different estimand from E and is reported beside it, never subtracted from it.

Models, training, and inference. θ_0 is Qwen3-4B-Base after supervised warm-up on PRM800K traces. All arms share every training setting (148 steps, batch 4, learning rate 10^{-6}, fp32 master weights and optimizer state, bf16 autocast); only the target changes. Per-assignment intervals come from a problem-clustered bootstrap (B=20,000); the primary interval for each arm is a t-interval across the three assignments (df $= 2$). An interval that contains zero is reported as not detected. A same-sampler placebo (re-drawing S and U after training, 1000 times) and the SFT arm, whose sign is known in advance, serve as controls.

3 Replacing the target weakens and can reverse the trace

Weaker in most settings, reversed in the main one. Two statements have different amounts of support and we keep them apart. Replacing the target lowers E in seven of the eight student settings we ran (Appendix A), which span a base-model student, a natural-text corpus, and a second model family; the exception is a Gemma-3 base student whose spread across assignments is too large to read. Lowering it below zero is established for three SFT students: two sizes of one family and one student of a second family. In the main configuration (Table 1) the same problems, assignments, and budget give $+0.0161$ under fine-tuning and -0.0192 under hard distillation from the RL-tuned teacher, a paired difference of -0.0353 within assignments. The student's own argmax carries no information beyond θ_0 and still produces -0.0128, so an external teacher is not needed for the

Table 2: Online contrast (trained rollout minus untrained sibling, nats/token) for four on-policy arms. Same student and recipe as Table 1; 592 pairs from 363 prompts, prompt-clustered bootstrap. The power gate is the pre-registered minimum detectable effect at 80% power, threshold 0.0065. The identity floor comes from an earlier round trained with int8 optimizer state and one seed.

Arm	Target	Online contrast [95% CI]	Power gate
Positive control	SFT, own tokens	$+0.0177$ $[+0.0168, +0.0186]$	met
Soft	top-20 reverse KL to T1	-0.0018 $[-0.0032, -0.0004]$	met
Hard, seed 1	argmax labels of T1	-0.0132 $[-0.0181, -0.0084]$	not met
Hard, seed 2	argmax labels of T1	-0.0137 $[-0.0184, -0.0090]$	not met
Identity floor (earlier round)	soft, teacher = student	-0.0012	

reversal. Soft targets are negative as well, at about a fifth of the hard-target size. Among the teachers we tested, the contrast is most negative for the student itself and for its RL-tuned derivative (self and T1). It is smaller in magnitude for the 8B teacher trained with the student's SFT recipe (-0.0024), and the effect with the raw 8B base teacher is not detected. In the second family, the student's own argmax gives a negative contrast, whereas the contrast with a public instruction-tuned teacher is not detected.

The reversal replicates. On an independent pool drawn from DeepMath the pattern repeats ($+0.0183$ against -0.0176; paired -0.0359). On the pool of our earlier experiments it repeats too ($+0.0221$ against -0.0215), and 1000 same-sampler placebo re-draws place all nine arm-level readings outside the null ($p < 0.001$). A 1.7B student trained by the same recipe gives a negative self-argmax arm (-0.0048 $[-0.0065, -0.0030]$) with SFT positive ($+0.0089$). One setting, a Gemma-3 base student, is too noisy to read either way. Extending the main hard arm to three passes makes E about six times more negative (-0.116 $[-0.144, -0.087]$, earlier precision round) while an identity-teacher control of the same length does not move, so the effect is not a transient of the first pass.

A second family replicates the reversal when only the family changes. On a Llama-3.2-3B student warmed up with the same recipe on its own pool, fine-tuning gives $E = +0.0346$ $[+0.0294, +0.0398]$ and hard distillation on the student's own argmax gives -0.0170 $[-0.0218, -0.0122]$, with every assignment negative on its own; this arm repeats the self-argmax protocol of Table 1 on the second family, where that arm reads -0.0128. With the public Llama-3.1-8B-Instruct model as teacher instead, the hard arm is not detected (-0.0046 $[-0.0118, +0.0026]$) although the paired difference hard$-$SFT remains negative (-0.0392 $[-0.0514, -0.0269]$, against -0.0353 in the first family); the identity-teacher control of this family reads -0.0029 $[-0.0038, -0.0020]$. The second family thus repeats the pattern of the first: the reversal is clear with the student itself as teacher and is not detected with a public instruction-tuned teacher (Appendix H).

On-policy resampling gives the same direction; we report it as preliminary. Four online arms, 148 steps each at full precision, use the online contrast of §2 (Table 2). The positive control, fine-tuning on the trained rollout's own tokens, reads $+0.0177$ $[+0.0168, +0.0186]$, so the online instrument registers a positive trace when one is expected. The soft target, top-20 reverse KL and the mainstream on-policy objective, reads -0.0018 $[-0.0032, -0.0004]$, negative as under fixed prefixes and of the same size as three seeds of an earlier low-precision round; it is about 1.5 times the identity floor measured in that earlier round, which we did not re-measure at full precision. The hard target reads -0.0132 $[-0.0181, -0.0084]$ and -0.0137 $[-0.0184, -0.0090]$ in two seeds, about 70% of the fixed-prefix hard arm. Its pre-registered power gate, a minimum detectable effect of 0.0065 at 80% power, was not met (0.0092 and 0.0082), so these two readings are reported without a verdict. No arm collapsed: in any 20-step window at most 3% of the control's rollouts and 12% of the hard arms' rollouts reached the length limit. An RL arm (DAPO on the included problems, one assignment, ten steps) shifts the likelihood of excluded trajectories about 700 times less than the distillation arms and reads $+0.00005$ $[-0.00002, +0.00013]$; we report it as a single uninformative point (Appendix J) and make no claim about RL objectives.

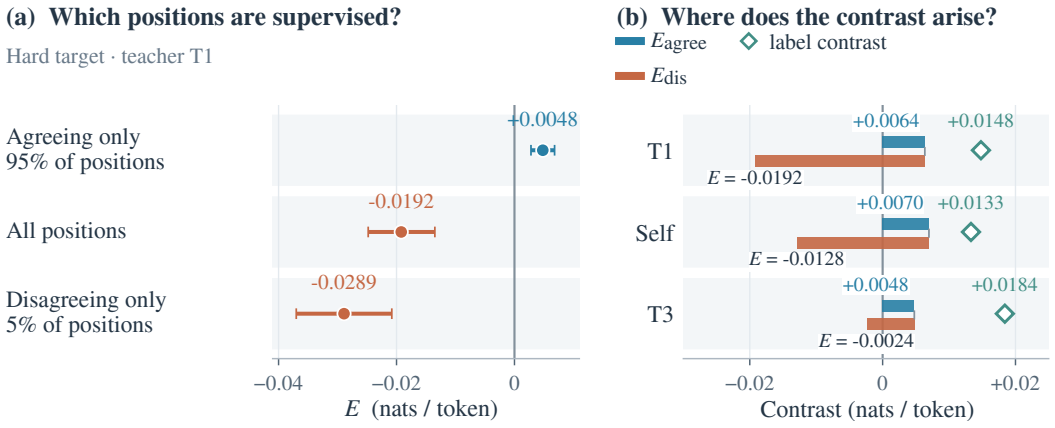


Figure 3: (a) E when supervision is restricted to agreeing positions, to disagreement positions, or applied to all positions (hard target, teacher T1). (b) Exact split of E into agreeing and disagreeing parts for three teachers: the blue bar runs from zero to E_{agree}, the orange bar starts where the blue bar ends and runs to $E = E_{\text{agree}} + E_{\text{dis}}$, and the hollow diamond is the label contrast measured on the same prefixes.

4 Where the negative sign comes from

Trajectory tokens and teacher labels move in opposite directions. If training leaves a trace on whatever the loss asks the model to predict, then the labels should become more likely among included problems even as the trajectories become less likely. We score, on the same teacher-forced prefixes, the log-probability of the label a_t in place of the token y_t and compute the same contrast. For all three teachers this label contrast is positive with intervals excluding zero (T1 $+0.0148$, self $+0.0133$, T3 $+0.0184$), while E on the trajectories is negative. For T3 the label contrast is eight times the size of the trajectory contrast and of opposite sign.

The sign is produced at disagreement positions. The teacher's label agrees with the trajectory's token at 94.8% of positions; we call the remaining 5.2% *disagreement positions*. Splitting E exactly into agreeing and disagreeing parts, $E = E_{\text{agree}} + E_{\text{dis}}$, gives $E_{\text{agree}} > 0$ and $E_{\text{dis}} < 0$ for every teacher (Figure 3b; Appendix B), and for the self-argmax arm of the second family ($+0.0135$ and -0.0305). Masking the loss tests whether the sign follows the position type. Supervising only disagreement positions yields $E = -0.0289$ $[-0.0370, -0.0208]$; supervising only agreeing positions yields $+0.0048$ $[+0.0028, +0.0068]$ (Figure 3a). Adding the disagreement positions also lowers the contrast at the agreeing positions, by -0.0036 $[-0.0039, -0.0032]$. A position where the student had written its token with probability 0.72 ends with that token at 10^{-9} and the teacher's token above 0.9999 (Appendix C); at agreeing positions the token is already likely and can gain at most a fraction of a nat, whereas at a disagreeing position it can lose twenty. A few such positions pull down a trajectory's mean log-probability, so after distillation the endpoint loss alone does not rank members above non-members (AUC 0.45–0.49, §5). Min-K% (Shi et al., 2024) scores a text by its least likely tokens; after distillation those are the disagreement positions of a member trajectory, so we expect it to reverse as well, though we did not test it. Concurrent work on on-policy distillation locates the task benefit at a few positions under selection rules of its own (Wang et al., 2026b; Liu et al., 2026; Xu et al., 2026), and shows that a fixed negative advantage on the student's lowest-probability tokens matches teacher-supervised distillation with no teacher at all (Ding & Zhang, 2026); position masking as a training intervention appears there as well (Wang et al., 2026b). What we measure after the intervention is the trace that inclusion leaves on the whole trajectory; the self-argmax arm of Table 1, plain cross-entropy on the student's own argmax, is the closest analogue in our design to that teacher-free result, although the two updates differ.

Table 3: Membership detection on held-out test problems (mean of three assignments). The direction "higher score ⇒ member" is fixed in advance. TPR is reported at the realized false-positive rate closest to 1% (0.94%); its resolution is $1/296 = 0.003$, so the three teacher values are one to three resolution steps above zero. The identity-teacher control has $E = -0.0011$, a seventeenth of the T1 arm's. Label-based scores use the teacher's labels; the SFT arm has no teacher, and they were not computed for the control.

Score	Needs	SFT	Identity	T1	self	T3
Reference increment, AUC	$+\theta_0$	**0.844**	**0.408**	**0.408**	**0.427**	**0.447**
Reference increment, TPR at FPR 0.94%	$+\theta_0$	0.27–0.37	0.014–0.024	0.007	0.005	0.010
Endpoint loss, AUC	final model	0.557	0.490	0.453	0.466	0.486
Label loss (calibrated), AUC	final model + labels	–	–	0.672	0.661	0.584
Label increment, AUC	$+\theta_0$ + labels	–	–	0.615	0.617	0.752

Why fine-tuning is protected to first order and distillation is not.

Lemma 1. *Fix N problems, draw m uniformly without replacement for training, and take one preconditioned step of size η on the mean training loss over the drawn problems, with a fixed positive semidefinite preconditioner P that does not depend on the draw. Let g_i be the gradient of the mean scored log-likelihood of problem i, averaged over its four trajectories, and h_i the gradient of its mean training loss, both at θ_0; the unit is the problem, as in the random assignment. Then*

$$\mathbb{E}[E] \;=\; -\frac{\eta N}{m(N-1)} \cdot \frac{1}{N} \sum_i (g_i - \bar{g})^\top P\,(h_i - \bar{h}) \;+\; O(\eta^2).$$

For fine-tuning on the scored text, $h_i = -g_i$, so the leading term is a non-negative quadratic form; the $O(\eta^2)$ remainder is not controlled here, so the statement is about the first-order term and not a bound for finite steps. Replacing the target breaks the identity $h_i = -g_i$; the sign is then set by the empirical covariance between what is scored and what is trained, and our measurements show that it can be negative. The lemma covers one step with a fixed preconditioner; it does not describe multi-step Adam, and it does not apply to the online contrast, where the scored rollout took part in the update. We use it to explain the direction and not to predict magnitudes (proof in Appendix K).

5 Implications for membership auditing

Reference-based scores read backwards after distillation and after the control. Four kinds of membership score are in play. An endpoint score uses only the final model; a reference score compares the final model with the initial student θ_0; a label score requires the teacher's labels; learned or black-box attacks fit a detector on labelled members of similar models (Ilić et al., 2026), and we do not test them. Our result concerns the fixed-direction reference score, the one that asks most directly whether a candidate became more likely. We follow the low-FPR regime that membership evaluations have converged on (Carlini et al., 2022) and report AUC as a summary; Appendix G gives intervals for every AUC in Table 3. The reference increment detects fine-tuning (TPR 0.27–0.37 at a realized FPR of 0.94%, AUC 0.84) and ranks members below non-members for all three hard-distillation teachers (AUC 0.41–0.45, every interval below 0.5; Table 3). The decisive row is the identity-teacher control. Its contrast is a seventeenth of the T1 arm's (-0.0011 against -0.0192), and the score reverses just as strongly (AUC 0.41; TPR 0.014–0.024 at the same operating point, within a few resolution steps of the false-positive rate, as are the three teachers' values of 0.005–0.010). Paired against the control, the T1 arm is indistinguishable ($+0.000$), the self arm is not detected ($+0.019$), and only the T3 arm is less reversed ($+0.039$, detected). The pattern holds on the earlier pool (SFT 0.93; hard 0.37; soft 0.39–0.40; identity 0.41) and in the second family (SFT 0.86; identity 0.39; hard 0.49, less reversed than the control by 0.098, detected). Every distillation arm we tested and the control reorder members and non-members in the wrong direction, and the ordering reflects the direction of the update rather than its size; the one RL arm, which barely moved the parameters, sits at chance (Appendix J). The reversal of the score is therefore not the reversal of E: the contrast is seventeen times larger for the T1 arm than for the control while the score is the same, and only the contrast separates them.

Endpoint-only scores read little; label scores recover the signal. Without the reference model the plain loss attack detects fine-tuning only weakly (AUC 0.56 $[0.51, 0.60]$) and is not detectably

different from chance for the control and the three teachers (0.45–0.49, every interval containing 0.5). Scoring the teacher's labels instead of the trajectory detects membership for every teacher. The best label score depends on the teacher (label loss 0.67 for T1, label increment 0.75 for T3), so we do not single one out.

What an auditor can and cannot conclude. Because the objective is unknown, a one-sided score is uninterpretable: a low reading is consistent with distillation, a high reading with fine-tuning, and either with drift. For data the auditor randomized into or out of training before release, the null of non-use makes S and U exchangeable, so a contrast of either sign that excludes zero shows that the data was used; among the objectives we tested, a positive contrast accompanied fitting the scored text and a negative one accompanied target replacement. The direction of a score should be set on independent calibration data with the objective known, and the calibrated discrimination reported alongside; calibration alone moves the reference score from 0.41 to 0.59 in our setting. When the teacher's labels are available, the label score should be compared with the trajectory score. The randomization has to happen before release: a data owner who withholds a random subset of a dataset before publishing it, or who plants a random subset of canaries, satisfies the assumption, whereas texts scraped from the web have no withheld counterpart and give the auditor no null distribution. The exchangeability argument also needs the trainer's pipeline to be blind to the assignment: filtering the release in a way that depends on which texts were withheld breaks it. A trainer who independently holds copies of some withheld texts does not break it, but then the test speaks only about this release, and withheld does not mean unseen. None of this applies to data obtained after the fact, and none of it establishes non-use: a null contrast is consistent with data that was never used, with an objective that leaves no trace at this power, and with stages of a pipeline whose traces cancel. Likelihood audits after post-training can show use, not non-use.

6 Related work

Memorization and membership under distillation. Students can leak their teacher's training data (Jagielski et al., 2023; Zhang et al., 2025c; Behrens & Zdeborová, 2026), distillation memorizes less than fine-tuning in extraction-based measures (Borkar et al., 2026), and distilled language models are not consistently less vulnerable to membership attacks than their teachers (Cui et al., 2025). These studies concern the teacher's data, extraction rates, or attack success on nominal training and held-out sets. We measure the signed effect of training on the student's own trajectories under randomized inclusion and find that its direction can reverse.

Learning dynamics and likelihood displacement. Preference optimization can lower the likelihood of preferred responses (Razin et al., 2025; Ren & Sutherland, 2025), and post-training objectives differ in what they change in a model (Chu et al., 2025; Yue et al., 2025). Repeated on-policy distillation on a single example keeps improving accuracy (Fu et al., 2026); in our runs the trained trajectories become less likely while task accuracy rises for the T1 teacher and does not detectably move for the others (Appendix E). We study target replacement on fixed prefixes and connect the resulting displacement to the reading of membership audits.

Reliability of membership inference. Membership inference on LLMs often performs near chance (Duan et al., 2024), published evaluations can be confounded by distribution shift (Das et al., 2025; Meeus et al., 2025), and membership scores cannot on their own prove training (Zhang et al., 2025a); in RL with verifiable rewards the signal moves to behaviour (Liu et al., 2025). Two observations come closest to ours. Several settings in Duan et al. (2024) have AUC well below 0.5, which that work discusses in terms of reference models and distribution differences rather than of the training objective, and brief GRPO training drives loss-based contamination detectors to chance (Wang et al., 2026a), which that work traces to importance sampling and clipping in the RL objective and to the assumption that training samples incur lower loss. Dataset inference (Maini et al., 2024) absorbs the direction of each feature into learned coefficients, which requires labelled members of the same model. Randomized inclusion follows counterfactual memorization and one-run privacy auditing (Feldman & Zhang, 2020; Zhang et al., 2023; Steinke et al., 2023). Our design removes distribution shift by construction and shows a failure of a different kind: the signal is present, and its sign depends on the objective.

Sparse supervision in on-policy distillation. A small fraction of positions carries the task benefit of on-policy distillation (Wang et al., 2026b; Liu et al., 2026; Xu et al., 2026; Ding & Zhang, 2026), and restricting supervision to positions where the student's argmax lies within the teacher's top-k is a workable selection rule (Huang et al., 2025), a looser notion of agreement than the exact match of teacher argmax and sampled token used here. These works choose which positions to supervise in order to improve training. We study what supervising those positions does to the trace left on the original trajectory and on the labels.

7 Limitations and conclusion

Limitations. *Online evidence.* The positive control and the soft target have one seed each; the soft effect is about 1.5 times an identity floor measured only with int8 optimizer state; the two hard-target seeds are negative but did not meet their pre-registered power gate; the online arms have no task evaluation. *Scope of the reversal.* The reversal is established for hard targets on three SFT students of two families, with the student itself or its RL-tuned derivative as teacher; teachers built from a different base model give smaller contrasts, and a public instruction-tuned teacher in the second family is not detected. Full-vocabulary soft distillation was not run, and the soft arms of the second family were excluded by a pre-registered drift gate. *Replication.* The three assignments are three draws of the included set from one pool under one training seed; the across-assignment interval reflects the draw and not optimizer randomness, teacher sampling, or an independently sampled pool, and the three draws share the pool. *Masking.* The masked arms keep the loss denominator fixed, so the disagreement-only arm receives gradient from about 5% of positions and the agreeing-only arm from about 95%; the two interventions show that the sign follows the position type, and we did not match total gradient scale between them. *Task evaluation.* Task accuracy was measured for the hard arms, the masked arms, and the four teachers of the main configuration; the SFT, soft-target, identity-control, online, and second-family arms were not evaluated on the task. *Theory.* The lemma covers one step with a fixed preconditioner and does not extend to multi-step Adam or to the online contrast. *Data overlap.* About 9% of pool problems overlap with data seen in warm-up or teacher training; S and U come from the same pool, so E is unbiased, but held-out does not mean never seen. *Access.* Label-based scores require the teacher or its labels. *RL.* One ten-step arm is uninformative and bounds nothing about longer RL runs. *Precision.* The fixed-prefix soft-target arms were trained with int8 optimizer state and were not rerun at fp32.

Conclusion. We asked whether training on a trajectory makes it more likely than a trajectory that was not trained on. Under randomized inclusion the answer depends on the training target. Fine-tuning gives the included trajectories the larger gain. Distilling the same prefixes onto a teacher's labels gives them the smaller gain, and in our main settings a negative contrast. That contrast is produced at the few positions where the labels disagree with the trajectory; the supervised labels, meanwhile, gain more likelihood on included than on excluded problems. Preliminary on-policy experiments also yield negative contrasts. A reference-based membership score reverses after every distillation arm on which we computed it and after the identity-teacher control, so a reversed score on its own says nothing about the objective. Under randomized inclusion, a contrast in either direction that excludes zero shows that the data was used; a null does not establish non-use. Two settings would sharpen the scope of the reversal: full-vocabulary soft distillation, and a sequence of teachers at controlled distances from the student within one family, since our three teachers differ in size, recipe, and lineage at once.

AI use statement

In this work we used generative AI tools to refine hypotheses and to give feedback on experimental design, to implement and debug training, scoring, and analysis code, to assist in writing the proof in Appendix K, and to help interpret results. The training trajectories were sampled from the student models under study, as part of the experimental design; beyond that, we did not use generative AI tools to generate synthetic data, to clean or reformat datasets, or to translate; qualitative or thematic data analysis is not applicable to this work. We additionally used generative AI tools to draft and edit parts of the text, to create scientific figures, to search for and check literature, and to format references. We have reviewed all AI-assisted work: every reported number was produced by our own versioned scripts from logged experiments, AI-assisted code was tested against pre-registered

acceptance checks before use, the proof was checked line by line by the authors, and all cited works were verified against their official sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility statement

All training and scoring settings are fixed across arms and listed in §2 and Appendix L. Every reported number is produced by versioned scripts from logged training commands, and the verdict rule for each arm was written down before the corresponding reading. Anonymized code for training, scoring, and the estimators, together with the per-trajectory scores behind every table, will be shared with reviewers through an anonymous repository once the discussion forum opens, and released publicly with the paper.

References

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

Freya Behrens and Lenka Zdeborová. Dataset distillation for memorized data: Soft labels can leak held-out teacher knowledge. In *International Conference on Learning Representations*, 2026. arXiv:2506.14457.

Jaydeep Borkar, Karan Chadha, Niloofar Mireshghallah, Yuchen Zhang, Irina-Elena Veliche, Archi Mitra, David A. Smith, Zheng Xu, and Diego Garcia-Olano. Memorization dynamics in knowledge distillation for language models. *arXiv preprint arXiv:2601.15394*, 2026.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy*, 2022.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations*, 2023.

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*, 2025.

Ziyao Cui, Minxing Zhang, and Jian Pei. On membership inference attacks in knowledge distillation. *arXiv preprint arXiv:2505.11837*, 2025.

Debeshee Das, Jie Zhang, and Florian Tramèr. Blind baselines beat membership inference attacks for foundation models. In *IEEE Security and Privacy Workshops (SPW)*, pp. 118–125, 2025.

Yi Ding and Ruqi Zhang. Does on-policy distillation really distill? from noisy teacher to self-improvement. *arXiv preprint arXiv:2608.31046*, 2026.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling*, 2024.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems*, 2020.

Zixuan Fu, Bingxiang He, Yuxin Zuo, Haohuan Huang, Jinqian Zhang, Ruhang Xiao, Cheng Qian, Qinyu Luo, Huan-ang Gao, Yudong Wang, Zhiyuan Liu, Ning Ding, and Chaojun Xiao. Rethinking on-policy distillation of large language models II: One training example. *arXiv preprint arXiv:2609.04172*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. arXiv:2306.08543.

Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. SelecTKD: Selective token-weighted knowledge distillation for LLMs. *arXiv preprint arXiv:2510.24021*, 2025.

David Ilić, Kostadin Cvejoski, David Stanojević, and Evgeny Grigorenko. Learning the signature of memorization in autoregressive language models. *arXiv preprint arXiv:2604.03199*, 2026.

Matthew Jagielski, Milad Nasr, Katherine Lee, Christopher A. Choquette-Choo, Nicholas Carlini, and Florian Tramèr. Students parrot their teachers: Membership inference on model distillation. In *Advances in Neural Information Processing Systems*, 2023.

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*, 2026.

Yule Liu, Heyi Zhang, Jinyi Zheng, Zhen Sun, Zifan Peng, Jiaheng Wei, Tianshuo Cong, Yilong Yang, and Xinlei He. Auditing data membership in reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2511.14045*, 2025.

Zhishuai Liu, Xingzi Xu, Mehmet Saygin Seyfioglu, Pan Xu, and Karim Bouyarmane. Extremely sparse supervision incentivizes reasoning ability. *arXiv preprint arXiv:2609.04565*, 2026.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. LLM dataset inference: Did you train on my dataset? In *Advances in Neural Information Processing Systems*, 2024.

Matthieu Meeus, Igor Shilov, Shubham Jain, Manuel Faysse, Marek Rei, and Yves-Alexandre de Montjoye. SoK: Membership inference attacks on LLMs are rushing nowhere (and how to fix it). In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2025. arXiv:2406.17975.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. In *International Conference on Learning Representations*, 2025.

Yi Ren and Danica J. Sutherland. Learning dynamics of LLM finetuning. In *International Conference on Learning Representations*, 2025.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, 2024.

Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. In *Advances in Neural Information Processing Systems*, 2023.

Han Wang, Haoyu Li, Brian Ko, and Huan Zhang. On the fragility of benchmark contamination detection in reasoning models. In *International Conference on Learning Representations*, 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.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026.

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*, 2025.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. In *Advances in Neural Information Processing Systems*, 2023.

Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramèr. Position: Membership inference attacks cannot prove that a model was trained on your data. In *IEEE Conference on Secure and Trustworthy Machine Learning*, pp. 333–345, 2025a.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Yang, and Hai Li. Min-K%++: Improved baseline for pre-training data detection from large language models. In *International Conference on Learning Representations*, 2025b. arXiv:2404.02936.

Ziqi Zhang, Ali Shahin Shamsabadi, Hanxiao Lu, Yifeng Cai, and Hamed Haddadi. Membership and memorization in LLM knowledge distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20074–20084, 2025c.

A ALL SETTINGS

Table 4: E for the SFT arm and the argmax arm, and their paired difference, in eight student settings; two students are listed with two teachers. Each row's argmax arm uses the teacher named in the row; "self" is the student's own argmax. Intervals are t-intervals over assignments where three assignments exist; "n.d." marks an interval containing zero. The main configuration is listed twice, once per hard-target teacher. A Gemma-3-4B base student (SFT $+0.0408$, argmax arm -0.0238, paired difference -0.0646) is omitted from the rows: its spread across assignments is about 29 times the estimate, so neither the sign of the arm nor that of the paired difference can be read; it is the one setting of the eight in which the paired difference is not detected.

Student	Argmax teacher	E_{SFT}	E_{argmax}	Paired difference
Qwen3-4B-SFT (main)	T1	$+0.0161$	-0.0192	-0.0353 $[-0.0427, -0.0278]$
Qwen3-4B-SFT (main)	self	$+0.0161$	-0.0128	-0.0289 $[-0.0336, -0.0243]$
Qwen3-4B-Base	self	$+0.0197$	$+0.0103$ n.d.	-0.0094
Qwen3-1.7B-SFT	self	$+0.0089$	-0.0048	-0.0137
Qwen3-1.7B-Base	self	$+0.0111$	-0.0052 n.d.	-0.0164
Qwen3-4B-SFT, natural text	self	$+0.0327$	-0.0481 n.d.	-0.0804
Gemma-3-4B-SFT	self	$+0.0463$	n.d.	-0.1059
Llama-3.2-3B-SFT	self	$+0.0346$	-0.0170	-0.0516 $[-0.0559, -0.0473]$
Llama-3.2-3B-SFT	Llama-3.1-8B-Instruct	$+0.0346$	-0.0046 n.d.	-0.0392 $[-0.0514, -0.0269]$

The Gemma-3-4B SFT student replicates the attenuation (paired difference -0.1059, detected) with each arm's own interval crossing zero, a pattern we attribute to the larger per-assignment spread of that family on this recipe rather than to a smaller effect. On the DeepMath pool (twelve arms, no interval crossing zero) the main configuration gives $+0.0183$ for SFT, -0.0176 for hard distillation from T1, and a paired difference of -0.0359. On the pool of our earlier experiments (439 items) it gives $+0.0221$ and -0.0215, and 1000 same-sampler placebo re-draws place all nine arm-level readings outside the null ($p < 0.001$). Extending the hard arm to three passes gives $E = -0.1155$ $[-0.1443, -0.0868]$ at 444 steps, against -0.0192 at 148, with no comparable change in an identity-teacher control of the same length.

B POSITION DECOMPOSITION

Table 5: Exact split of E into agreeing and disagreeing positions, and the label contrast, for the hard target with three teachers of the main configuration and the self-argmax arm of the second family (Llama). Agreement rate on the training pool: 94.82% of 672,476 supervised positions for T1, 94.7% for Llama self. The label contrast was not computed for the Llama arm.

Teacher	E	E_{agree}	E_{dis}	Label contrast
T1	-0.0192	$+0.0064$ $[+0.0040, +0.0087]$	-0.0255 $[-0.0331, -0.0179]$	$+0.0148$ $[+0.0095, +0.0202]$
self	-0.0128	$+0.0070$	-0.0198	$+0.0133$ $[+0.0097, +0.0169]$
T3	-0.0024	$+0.0048$	-0.0071	$+0.0184$ $[+0.0156, +0.0211]$
Llama self	-0.0170	$+0.0135$ $[+0.0098, +0.0173]$	-0.0305 $[-0.0346, -0.0264]$	–

The masked arms of Figure 3 keep the loss denominator fixed, so the masked losses are not rescaled to the number of supervised positions; the two masked contrasts need not add to the full one (-0.0241 against -0.0192).

C TWO DISAGREEMENT POSITIONS

Table 6: Two disagreement positions from one hard-distillation run (T1, first assignment), selected by a fixed rule: the most negative $\Delta \log p(y_t)$ among positions with $a_t \neq y_t$ and $\Delta \log p(a_t) > 0$, one per trajectory. They illustrate the mechanism, not its frequency.

Context	$y_t \to a_t$	$p_{\theta_0}(y_t)$	$p_\theta(y_t)$	$p_{\theta_0}(a_t)$	$p_\theta(a_t)$
`... - 115yz + z^2 [=] 0.`	" =" → " +"	0.72	1.3×10^{-9}	0.27	> 0.9999
`... all multiples of 3 [and] therefore equal`	" and" → ", "	0.084	4.8×10^{-11}	0.90	> 0.9999

D PER-ASSIGNMENT READINGS OF THE MAIN ARMS

Table 7: Δ_S, Δ_U, and E for each of the three assignments of the main configuration, with the per-assignment problem-clustered bootstrap interval for E. The drift Δ_U is several times larger than E in every hard arm.

Arm	Assignment	Δ_S	Δ_U	E	95% CI for E
SFT	1 / 2 / 3			$+0.0167$ / $+0.0152$ / $+0.0164$	
Hard, T1	1	-0.1049	-0.0834	-0.0214	$[-0.0282, -0.0148]$
Hard, T1	2	-0.0960	-0.0791	-0.0169	$[-0.0235, -0.0103]$
Hard, T1	3	-0.1013	-0.0822	-0.0191	$[-0.0258, -0.0126]$
Hard, self	1	-0.0706	-0.0572	-0.0134	$[-0.0189, -0.0082]$
Hard, self	2	-0.0676	-0.0561	-0.0115	$[-0.0169, -0.0063]$
Hard, self	3	-0.0714	-0.0578	-0.0136	$[-0.0190, -0.0083]$
Hard, 8B base (T2)	1	-0.0595	-0.0549	-0.0046	$[-0.0095, +0.0002]$
Hard, 8B base (T2)	2	-0.0593	-0.0586	-0.0006	$[-0.0052, +0.0039]$
Hard, 8B base (T2)	3	-0.0594	-0.0557	-0.0037	$[-0.0087, +0.0011]$

The T2 arm uses the raw Qwen3-8B base model as teacher; its MATH-500 accuracy after training is 0.001 with about 40% of answers hitting the length limit, so we treat it as a degenerate arm and do not include it in the main text. Its per-assignment training loss falls by 11–54%, so the arm did train; the contrast is small and not detected in any assignment.

E TASK ACCURACY

MATH-500 accuracy (greedy, 500 problems) of the initial student is 0.406. After the hard arm with teacher T1 it is 0.477; after supervising only agreeing positions 0.444; after supervising only disagreement positions 0.476. Two comparisons are involved. Relative to the initial student, the T1 arm gains about 7 points and the gain is readable per problem (below). Between training arms, every paired difference among the full, agreeing-only, and disagreeing-only arms has an interval that crosses zero (half-widths of 5–7 points), so the position experiments do not detect where the task benefit lives, only where the likelihood trace lives. Among the three teachers, T1 is the only one with a detected per-problem gain (4.8–9.4 points, $p \leq 0.02$); the self and T3 arms do not detectably move accuracy, and the T2 arm collapses. The contrast E, the drift Δ_U, and the label contrast are therefore not diagnostics of whether an arm learned the task: the self and T2 arms have the same drift (-0.057 and -0.056) with accuracies of 0.41 and 0.00.

F TOKEN CLASSES AND EXTREME POSITIONS

These readings come from the soft-target arm (top-20 reverse KL to T1) on the pool of our earlier experiments (439 items; $E = -0.0041$ $[-0.0054, -0.0029]$; the hard arm on that pool reads -0.0215, Appendix A) and were pre-registered there.

Table 8: Left: E restricted to token classes (earlier pool, soft target). Right: E after removing the tokens with the largest signed student/teacher excess log-ratio, the tokens the reverse-KL gradient pushes down hardest.

Class	Tokens	E	Removed	E on remainder	95% CI
Special	439	-0.0001	none	-0.0041	$[-0.0054, -0.0029]$
Punctuation	69,360	-0.0017	top 1%	-0.0024	$[-0.0030, -0.0018]$
Digit	46,022	-0.0073	top 5%	$+0.0005$	$[+0.0000, +0.0011]$
Content	195,545	-0.0036			

The effect is largest on digit tokens and not detectable on punctuation or special tokens, so it is not a formatting artifact. Removing the most extreme 1% of tokens leaves the contrast negative and reduces it by 42%; removing the most extreme 5% removes it. The negative sign is therefore carried by a substantial minority of tokens rather than by a handful of outliers or by every token alike, which is the same picture as the disagreement-position split of Appendix B on the main pool. A partition by the baseline probability of each token was also pre-registered; it produces an artifact of the log scale (a low-probability token can lose tens of nats while a high-probability token can gain at most a fraction of one) and we do not interpret it.

G AUDIT SCORES, CALIBRATION, AND THE NULL HYPOTHESIS

Scores. For trajectory i with tokens y_i and teacher labels a_i on the same prefixes, with $\bar{\ell}_\theta$ the mean per-token log-probability under model θ: the *endpoint loss* is $\bar{\ell}_\theta(y_i)$ (we use mean log-probability rather than negative log-likelihood throughout, so that "higher" means member for every score); the *reference increment* is $d_i = \bar{\ell}_\theta(y_i) - \bar{\ell}_{\theta_0}(y_i)$; the *label loss* is $\bar{\ell}_\theta(a_i)$; the *label increment* is $g_i = \bar{\ell}_\theta(a_i) - \bar{\ell}_{\theta_0}(a_i)$. Each score is used as a membership statistic with the direction "higher $\Rightarrow$ member" fixed before any reading, without the $\max(A, 1-A)$ convention; an AUC below 0.5 therefore means that members rank below non-members under the assumed direction.

Split and calibration. The 1000 problems are split in half by problem with a fixed seed (20260816), preserving the S/U proportion in each half. All reported AUC and TPR values are computed on the test half; the calibration half is used only by the calibrated scores, where the sign of the score is chosen so that the calibration-half members score higher. The calibrated label loss in Table 3 is calibrated in this sense; the reference increment is reported uncalibrated and, in §5, after the same calibration (0.41 to 0.59). Intervals are problem-clustered stratified bootstraps (B=20,000) over the test half; paired differences between arms subtract AUCs within an assignment before the t-interval over assignments. AUC and TPR are computed over trajectories (296 members and 1704 non-members on the test half); the split and the bootstrap are by problem. TPR is reported at the realized false-positive rate closest to 1% (0.94%; the resolution is $1/n_U$ on the test half).

Intervals for Table 3. Table 9 gives, for every AUC in Table 3, the t-interval across the three assignments (df $= 2$), the same construction as the primary interval of Table 1. Every reference-increment interval of a distillation arm and of the control lies below 0.5; every endpoint-loss interval of those arms contains 0.5.

Table 9: AUC with t-intervals across the three assignments (df $= 2$) for the scores of Table 3. Label-based scores use the teacher's labels; the SFT arm has no teacher, and they were not computed for the control.

Arm	Reference increment	Endpoint loss	Label loss (calibrated)	Label increment
SFT	0.844 $[0.804, 0.883]$	0.557 $[0.512, 0.601]$	–	–
Identity control	0.408 $[0.396, 0.419]$	0.490 $[0.440, 0.541]$	–	–
Hard, T1	0.408 $[0.378, 0.438]$	0.453 $[0.407, 0.500]$	0.672 $[0.609, 0.735]$	0.615 $[0.574, 0.657]$
Hard, self	0.427 $[0.385, 0.468]$	0.466 $[0.415, 0.518]$	0.661 $[0.600, 0.723]$	0.617 $[0.570, 0.664]$
Hard, T3	0.447 $[0.438, 0.456]$	0.486 $[0.436, 0.536]$	0.584 $[0.535, 0.634]$	0.752 $[0.712, 0.792]$

Null hypothesis of the randomized test. The data owner draws a fixed-size random subset of problems as S before release and withholds the rest as U; in our runs $|S| = 148$ of $N = 1000$.

Under the null that the released data did not enter training, the whole training and scoring pipeline is independent of the assignment, so the labels S/U are exchangeable with respect to every score, and a permutation test that re-draws a subset of the same size by problem has exact level. We implement it as the same-sampler placebo: the assignment is re-drawn 1000 times with the trained model fixed, and the observed contrast (or AUC) is compared with this null distribution. A rejection in either direction shows that the assignment mattered to training, which is what "the data was used" means here; it does not identify at which stage or through which objective, and it says nothing about data the owner did not randomize. On the main pool, the label-permutation null for the reference-increment AUC is centred on 0.5 for both the hard arm and the control, and the observed values (0.41) lie outside it.

H SECOND FAMILY

Llama-3.2-3B-Base student warmed up with the student recipe on PRM800K (Appendix L); exposure pool of 1000 problems with four trajectories each, sampled from this student; the three assignments use the same problem sets as the main configuration; same estimator and intervals. SFT $+0.0346$ $[+0.0294, +0.0398]$; hard target from the student's own argmax -0.0170 $[-0.0218, -0.0122]$, with per-assignment readings -0.0148 $[-0.0204, -0.0094]$, -0.0186 $[-0.0251, -0.0123]$, and -0.0175 $[-0.0231, -0.0121]$; hard target from Llama-3.1-8B-Instruct -0.0046 $[-0.0118, +0.0026]$; identity-teacher control -0.0029 $[-0.0038, -0.0020]$; paired difference Instruct-hard$-$SFT -0.0392 $[-0.0514, -0.0269]$. Within this family, the self-argmax arm was run after the other three with the training command of the Instruct-teacher arm and only the teacher path changed. The soft arm of this family had a drift $|\Delta_U|$ of 0.164, 17.5 times the Qwen value, and only one assignment; it was excluded by the pre-registered variance gate. Reference-increment AUC: SFT 0.857 (TPR 0.54–0.62 at FPR 0.94%), hard 0.485, control 0.387; hard minus control $+0.098$, detected. Splitting the self-argmax contrast by position type gives $E_{\text{agree}} = +0.0140, +0.0118, +0.0147$ and $E_{\text{dis}} = -0.0289, -0.0304, -0.0322$ in the three assignments (t-intervals across assignments $[+0.0098, +0.0173]$ and $[-0.0346, -0.0264]$); in each assignment the two parts add to E within 10^{-9}.

I ONLINE ARMS

Each of the four arms trains for 148 steps at batch 4 with fp32 master weights and optimizer state. At every step two rollouts are drawn per prompt; a coin picks the one to train on; both are scored by the final model and by the generating model, and the paired difference of the two changes is averaged over 592 pairs from 363 prompt clusters with a clustered bootstrap (B=20,000). Arms: positive control (SFT on the trained rollout), soft (top-20 reverse KL to T1), hard seed 1 and hard seed 2 (argmax labels from T1); the soft and hard-seed-1 arms share generation and coin seeds with the control. Readings: $+0.0177$ $[+0.0168, +0.0186]$; -0.0018 $[-0.0032, -0.0004]$; -0.0132 $[-0.0181, -0.0084]$; -0.0137 $[-0.0184, -0.0090]$. Minimum detectable effects at 80% power: 0.0015, 0.0025, 0.0092, 0.0082; the pre-registered threshold was 0.0065. Pooling the two hard seeds (1184 pairs) was not pre-registered and is reported without a verdict: the pooled contrast is -0.0134, with interval $[-0.0167, -0.0102]$ when the bootstrap clusters by prompt (363 clusters) and $[-0.0169, -0.0101]$ when it clusters by arm and prompt (726 clusters); the corresponding minimum detectable effects are 0.0087 and 0.0062, so the pooled reading meets the threshold only under the less conservative clustering. Collapse checks are summarized below; the thresholds were fixed before the run.

Check	Observed (all four arms)	Threshold
Rollouts at the length limit, worst 20-step window	1.8%–6.1% (hard arms up to 12%)	30%
Final / initial mean rollout length	0.82–1.15	2.0

An earlier low-precision round of the soft arm gave -0.0022, -0.0019, -0.0023 in three seeds, and an identity-teacher online control in that round gave -0.0012.

J REINFORCEMENT LEARNING ARM

DAPO with verifiable rewards (the recipe used to produce T1: learning rate 3×10^{-6}, group size 8, rollout batch 32, no KL penalty) on the 148 included problems of the first assignment, ten steps, from θ_0. $E = +0.00005$ $[-0.00002, +0.00013]$, minimum detectable effect 0.0001; $\Delta_S = +0.00017$, $\Delta_U = +0.00012$; reference-increment AUC 0.519 $[0.482, 0.556]$; MATH-500 $0.404 \to 0.396$. The drift $|\Delta_U|$ is about 1/700 of that of the hard-distillation arms, so the arm bounds nothing about longer RL runs.

K PROOF OF LEMMA 1

The lemma asks a simple question: after one small training step on a random subset S of the problems, how much more does the likelihood of an included problem change than the likelihood of an excluded one, on average over which subset was drawn? The proof has three steps: write the one-step likelihood change of each problem as a sum of pairwise terms, average those terms over the random draw of S, and recognize the result as a covariance.

Step 1: the effect of one step on each problem. Let $A_{ij} = g_i^\top P\, h_j$ measure how much a training step taken for problem j moves the scored log-likelihood of problem i; A_{ii} is the effect of a problem on itself. One preconditioned step on the mean training loss over S moves the parameters by $\theta - \theta_0 = -(\eta/m)\, P \sum_{j \in S} h_j$. To first order in η the scored log-likelihood of problem i changes by

$$d_i \;=\; g_i^\top(\theta - \theta_0) \;=\; -\frac{\eta}{m} \sum_{j \in S} A_{ij},$$

so the contrast for a particular draw of S is

$$E \;=\; \frac{1}{m}\sum_{i \in S} d_i - \frac{1}{N-m}\sum_{i \notin S} d_i \;=\; -\frac{\eta}{m}\Big[\frac{1}{m}\sum_{i \in S}\sum_{j \in S} A_{ij} - \frac{1}{N-m}\sum_{i \notin S}\sum_{j \in S} A_{ij}\Big] + O(\eta^2).$$

The first bracketed sum collects the influence of the training set on its own members; the second collects its influence on the excluded problems.

Step 2: average over the random draw. S is a uniform subset of size m, so $\Pr(i \in S) = m/N$, $\Pr(i \in S, j \in S) = m(m-1)/(N(N-1))$ for $i \neq j$, and $\Pr(i \notin S, j \in S) = m(N-m)/(N(N-1))$. The diagonal terms A_{ii} appear only in the first sum, since a problem cannot be both included and excluded. Taking the expectation of each sum term by term,

$$\begin{aligned}\mathbb{E}[E] &= -\frac{\eta}{m}\Big[\frac{1}{N}\sum_i A_{ii} + \frac{m-1}{N(N-1)}\sum_{i \neq j} A_{ij} - \frac{m}{N(N-1)}\sum_{i \neq j} A_{ij}\Big] + O(\eta^2) \\ &= -\frac{\eta}{m}\Big[\frac{1}{N}\sum_i A_{ii} - \frac{1}{N(N-1)}\sum_{i \neq j} A_{ij}\Big] + O(\eta^2).\end{aligned}$$

The off-diagonal terms nearly cancel between the two sums: what remains is the average self-influence A_{ii} minus the average cross-influence A_{ij}.

Step 3: rewrite as a covariance. Using $\sum_{i \neq j} A_{ij} = \sum_{i,j} A_{ij} - \sum_i A_{ii}$, the bracket equals

$$\frac{1}{N-1}\Big[\sum_i A_{ii} - \frac{1}{N}\sum_{i,j} A_{ij}\Big] \;=\; \frac{1}{N-1}\sum_i (g_i - \bar{g})^\top P\,(h_i - \bar{h}),$$

where $\bar{g}$ and $\bar{h}$ are the means over problems. Substituting gives the form stated in the lemma. When the training loss is the negative of the scored log-likelihood, $h_i = -g_i$, the sum becomes $-\sum_i (g_i - \bar{g})^\top P\,(g_i - \bar{g})$, which is non-positive because P is positive semidefinite, so the leading term of $\mathbb{E}[E]$ is non-negative. When the training target is replaced, h_i and g_i are gradients of different quantities, the sum has no fixed sign, and $\mathbb{E}[E]$ can be negative. □

L HYPERPARAMETERS AND DATA OVERLAP

Student warm-up: Qwen3-4B-Base, full-parameter SFT on PRM800K traces, learning rate 10^{-5}, one epoch (1837 steps). Teacher T1: θ_0 plus 100 steps of DAPO on MATH-train (7498 problems). Teacher T3: Qwen3-8B-Base with the same SFT recipe. Pool: 1000 problems drawn from DAPO-Math-17k, four trajectories per problem sampled from θ_0. Training arms: 148 steps, batch 4, micro-batch 1, learning rate 10^{-6} with 10 warm-up steps, seed 42, fp32 master weights and optimizer state with CPU offload, bf16 autocast, teacher in bf16. Soft-target arms: at each supervised position the support is the top-20 token set of the student's current distribution, recomputed at every step; student and teacher probabilities are both renormalized on that support before the KL is taken, so teacher mass outside the support does not enter the loss (it is logged); the teacher's logits carry no gradient, the selection of the support is an index operation through which no gradient flows, and no temperature is applied; the forward-KL and reverse-KL arms share this implementation and differ only in the direction of the KL. Second-family student: Llama-3.2-3B-Base with the same warm-up recipe, end-of-sequence token 128001, same training settings. Scoring: mean per-token log-probability over the trajectory under teacher-forced prefixes, by θ_0 and by the trained model. Intervals: per-assignment problem-clustered bootstrap with B=20,000; across assignments, $t_{0.975,2} = 4.30$. Overlap: 42 pool problems appear in the warm-up data and 73 in the RL data of T1 (88 in the union), 8.8% of the pool. The identity-teacher copy is bit-identical to the student's weights at bf16; its argmax differs from the student's at 0.24% of positions, measured on the first assignment of the main configuration.