FROM CORRECT ANSWERS TO CORRECT SUPERVISION IN SELF-REWARDED TEST-TIME LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-rewarded learning reinforces the answers that win a model's votes. A correct response can therefore be available without becoming useful supervision. We study this selection step through recorded reward groups, controlled oracle-reward mixtures, and diagnostics of adaptation outcomes. Across 74 self-rewarded Llama-3-8B runs on screened GSM8K, a correct answer appears in 91.5% of training groups but is selected in 63.7%. We derive the group-dependent oracle weight at which an available correct response acquires a positive centered reward; its median is $1/3$ among wrong-vote groups containing that response. Across 288 reward-intervention runs, oracle anchoring raises mean held-out gain and reduces its variability. Selection correctness also provides information before the trajectory is complete: using the first 19 of 38 recorded groups, it reduces leave-one-seed-out gain-prediction error by 19.2% beyond response accuracy, with a 16.1% reduction in the reward-mixture cohort after controlling reward weight. These results distinguish producing a correct answer from selecting it for reinforcement, and establish reward selection as a measurable link between available responses and adaptation outcomes.

1 INTRODUCTION

Self-rewarded test-time learning uses a model's own answers to decide what it should reinforce. In majority-rewarded TTRL, sampled responses that agree with the plurality receive reward (Zuo et al., 2025). The procedure supplies a training target without an external verifier, but it also introduces a selection step between generating an answer and learning from it. A correct answer can be present in a rollout group while an incorrect answer wins the vote. In that case, the learning signal favors the selected answer and assigns the correct response a negative centered reward.

This distinction matters for understanding adaptation. Response accuracy measures how often the policy produces a correct answer; vote concentration measures how decisively responses agree. Neither quantity directly states which answer becomes the reward target. Existing work has shown that self-generated rewards can amplify errors, tracked the disappearance of correct answers during TTRL, and developed confidence-based or selective pseudo-labeling methods (Yuan et al., 2024; Lin et al., 2026; Wang et al., 2026; Yu et al., 2026; Yan et al., 2026). We ask how much information the selection step contributes to adaptation outcomes, and how changing its reward incentives alters those outcomes.

Our analysis separates three quantities in each sampled group: whether a correct answer is available, the fraction of correct responses, and whether the selected plurality is correct. The distinction is empirically substantial. In 2,812 groups from 74 majority-rewarded runs, correct answers are present in 91.5% of groups and selected in 63.7%. The remaining 27.8% contain correct responses that receive negative centered majority rewards. This measures a specific source of adverse supervision within groups that already contain a usable answer.

We connect this measurement to a controlled intervention:

$$r_\lambda = \lambda r_{\text{oracle}} + (1-\lambda) r_{\text{majority}}, \qquad 0 \le \lambda \le 1. \tag{1}$$

For a wrong plurality with frequency s and a correct-answer frequency $p > 0$, the correct response's centered reward becomes positive when $\lambda > s/(1-p+s)$. This threshold depends on the observed

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.

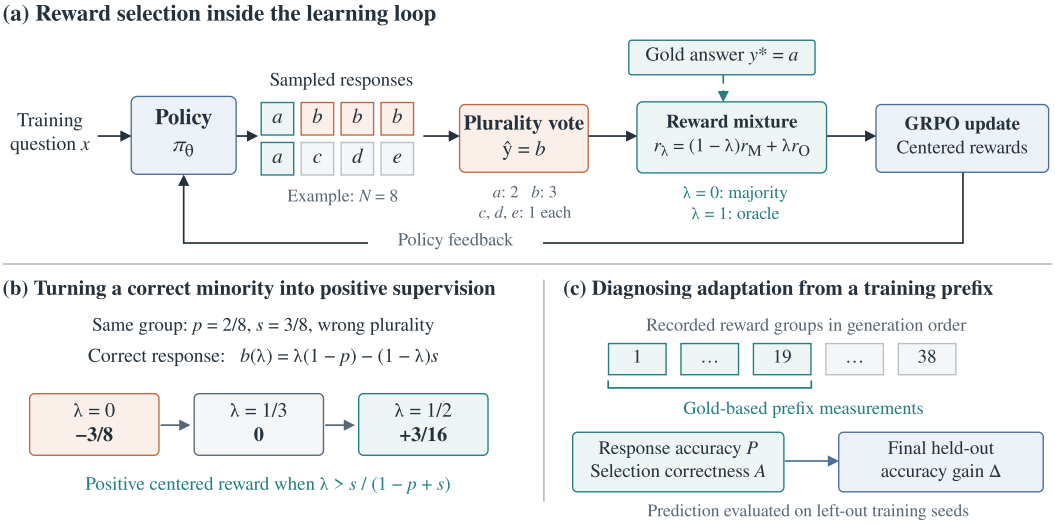

Figure 1: Reward selection connects sampled answers to supervision. (a) A policy samples eight responses, selects a plurality and learns from the mixture of majority reward r_M and oracle reward r_O. Gold labels enter the controlled oracle component and the diagnostic analysis; the majority objective uses the selected answer. (b) In the illustrated group, the correct answer a receives positive centered reward above $\lambda = 1/3$, while its reward exceeds the wrong plurality's above $\lambda = 1/2$. (c) Prefix diagnostics measure response accuracy and selection correctness from the same recorded groups and predict final gain on left-out training seeds.

group and can be below the $1/2$ reward-ordering crossover. The fixed-group analysis identifies which correct responses an oracle contribution can turn into positive supervision. The training experiment then measures the response to that intervention with refreshed groups and shared model parameters.

The evidence has two complementary parts. Across 288 matched-configuration runs, increasing the oracle weight raises mean held-out gains and concentrates their distribution. Within self-rewarded runs, selection correctness provides incremental diagnostic information even when measured from a prefix of the recorded trajectory. With the first 19 groups, adding vote correctness to response accuracy reduces final-gain prediction RMSE from 7.98 to 6.45 percentage points. The same comparison reduces error from 7.14 to 5.99 points in mixture runs after controlling reward weight. Quadratic accuracy controls and repeated seed-grouped splits preserve this pattern.

Figure 1 separates the learning loop, the group-level reward transformation, and the diagnostic readout. These results establish reward selection as both an intervention target and a diagnostic of adaptation. The contribution is a quantitative account of how available correct responses become reinforcement signals: a group-level threshold, its measured distribution, and a prefix diagnostic that adds information beyond response accuracy. Cross-model cohorts describe policy variation; supplementary sampling, input-edit and update-admission studies extend the measurement to related settings.

2 Related Work

Reward construction and reliable pseudo-labels. TTRL (Zuo et al., 2025) and self-rewarding language models (Yuan et al., 2024) learn from model-generated supervision. SCOPE combines confidence and subgroup consensus (Wang et al., 2026); denoising (Yu et al., 2026), selective and complementary pseudo-labels (Yan et al., 2026), and tool verification (Liao et al., 2026) improve the training targets through different information sources. Our oracle mixture is an experimental control on target incentives. Holding the learning configuration fixed exposes the response to correct supervision.

Correct-answer dynamics and adaptation diagnostics. Lin et al. (2026) track correct-answer extinction and label migration and introduce online safeguards. We share the focus on which answers receive reinforcement, and study the selection step through groupwise reward algebra and cross-run prediction. The diagnostic comparisons explicitly control response accuracy and restrict observations to prefixes of training. This tests whether correct plurality selection carries information beyond the supply of correct responses. Work on RLVR sampling efficiency (Yue et al., 2025) and TTRL input sensitivity (Khattar et al., 2026) motivates the supplementary studies. Additional connections appear in section A.

3 Method: Measuring Reward Selection

3.1 Reward construction

Let $y^*(x)$ be the gold answer and $\hat{y}_N(x)$ the plurality of N parsed responses. We use the term majority reward for $r_{\text{majority}}(a) = \mathbf{1}[a = \hat{y}_N]$, including groups without an absolute majority, and define $r_{\text{oracle}}(a) = \mathbf{1}[a = y^*]$. An unparseable response receives zero reward. The mixture in (1) changes the target incentives within each sampled group.

3.2 Availability, response accuracy and selection

For a sampled group g of N responses, let p_g be the fraction matching the gold answer, $a_g = \mathbf{1}[\hat{y}_g = y_g^*]$ indicate whether the selected plurality is correct, and s_g be the fraction matching that plurality. Unparseable responses contribute to N and receive zero reward. For a set of recorded groups $\mathcal{G}$, define

$$P = \frac{1}{|\mathcal{G}|}\sum_g p_g, \qquad H = \frac{1}{|\mathcal{G}|}\sum_g \mathbf{1}[p_g > 0], \qquad A = \frac{1}{|\mathcal{G}|}\sum_g a_g. \tag{2}$$

Here P measures response accuracy, H measures correct-answer availability, and A measures selection correctness, which we also call the vote audit. Their difference $H - A$ is the fraction of groups containing a correct answer that is not selected. For example, two correct responses among eight give $p_g = 1/4$ whether the remaining six responses are all different or include three copies of a single wrong answer. The correct answer wins the first vote and loses the second. This changes its majority reward from one to zero at the same response accuracy. The diagnostic comparison tests whether this distinction remains informative across observed training trajectories. These diagnostics use gold labels during analysis; majority-reward training uses the selected answer alone.

The reward callback stores answer counts, the chosen plurality and the gold answer in generation order. Each majority or mixture run contains 38 groups. We evaluate prefixes of 10 and 19 groups as well as all 38 groups, applying the same prefix to every feature in a comparison. Thus a diagnostic can be computed before the entire recorded trajectory is available.

3.3 When an available answer receives positive supervision

Centering rewards within a group makes the role of selection explicit. For a group with a wrong plurality and $p_g > 0$, a correct response has reward λ, whereas the group mean reward is $\lambda p_g + (1 - \lambda)s_g$. Its centered reward is therefore

$$b_g(\lambda) = \lambda(1 - p_g) - (1 - \lambda)s_g. \tag{3}$$

Proposition 1 (Correct-response reward threshold)**.** *In a fixed group with a wrong plurality and an available correct answer, $b_g(\lambda) > 0$ if and only if*

$$\lambda > \lambda_g^* = \frac{s_g}{1 - p_g + s_g} \leq \frac{1}{2}. \tag{4}$$

The inequality is strict when some responses match neither answer. At $\lambda = 1/2$, every available correct response in such a group has nonnegative centered reward. Positive standard-deviation normalization preserves its sign.

The result follows by solving (3) for λ and using $p_g + s_g \leq 1$. With correct plurality, the centered reward of a correct response is $1 - p_g$, independent of the mixture weight. Thus oracle anchoring

changes the sign of supervision specifically where available correct answers lose the vote. This is an identity for the sampled reward groups; the training study measures the resulting parameter updates, including normalization, clipping, the KL term and subsequent resampling. Additional fixed-target optimization results appear in section B.

Two thresholds describe different changes in supervision. Figure 1b illustrates a group with two correct responses, three copies of one wrong answer, and three other responses. Its threshold is $\lambda_g^* = 1/3$. The correct response's centered reward is $-3/8$ under majority reward, zero at $\lambda = 1/3$, and $+3/16$ at $\lambda = 1/2$. The sign change occurs because centering compares a response's reward with the mean over all eight responses, including the three receiving zero reward. At $\lambda = 1/2$, both the gold and wrong plurality answers receive reward $1/2$, above the group mean $5/16$.

The reward difference between the gold answer and the wrong plurality is $2\lambda - 1$, so their relative ordering switches at $1/2$. The threshold λ_g^* describes when correct responses contribute positive centered supervision; the ordering threshold describes when gold receives the larger reward. The distribution of the remaining answers determines the separation between these thresholds. This distinction makes partial oracle anchoring interpretable at the level of an observed group, before measuring the complete training response.

Signed supervision on correct responses. The group mean of centered reward carried by correct responses has the identity

$$S_g(\lambda) = p_g\big[\lambda(1 - p_g) + (1 - \lambda)(a_g - s_g)\big]. \tag{5}$$

It is the empirical covariance between response correctness and the mixture reward. If no correct response is sampled, $S_g = 0$ at every weight. If the correct response wins, $S_g = p_g(1 - p_g)$. If it loses, $S_g = p_g b_g(\lambda)$, and the threshold in (4) controls its sign. This separates three situations that a run's average reward cannot distinguish: absent correct responses, correct targets already selected, and available correct responses assigned adverse reward. The identity concerns sampled supervision; sequence probabilities and shared parameters determine how it translates into policy changes.

4 Experimental Setup

Screening, training and evaluation. The main study uses GSM8K (Cobbe et al., 2021). Screening draws four responses per problem with a direct-answer prompt and a 256-token cap, retaining problems with one to three correct responses. A fixed split assigns 200 such problems to training and 100 to evaluation. This in-band population isolates adaptation on problems that the base policy sometimes solves.

Training uses TRL GRPO (Shao et al., 2024), groups of eight, 300 optimization steps, a 0.02 KL coefficient, and QLoRA adapters (Hu et al., 2022; Dettmers et al., 2023). We compare majority, oracle and three intermediate rewards. The base and adapted policies are evaluated with the same evaluation seeds and batching. For each problem, the reported pass@k is the success indicator among the first k sampled responses, averaged over the 100 problems. Endpoint evaluation uses four samples per problem; a separate coverage evaluation uses 32 samples. Matching evaluation seeds reduces one source of sampling variability, although responses can differ across policies. Full configuration and metric definitions appear in section C.

Generation groups and optimization steps. The generation batch contains eight responses to one question, while each optimizer minibatch contains one response and gradient accumulation is one. A sampled group therefore supplies eight minibatches. The 300-step run produces 38 reward groups, with the final group used partially. The diagnostics weight generated groups equally. Prefixes of 10 and 19 groups use 80 and 152 sampled responses, respectively, from the same training process; they change the observation budget rather than the training schedule.

Diagnostic comparisons. All prediction models are ordinary least squares with an intercept, evaluated by leaving out every record sharing a training seed. We compare response accuracy P, its quadratic expansion, and P together with five concentration signals: margin, plurality share, distinct-answer count, entropy and identical-answer-group fraction. Each augmented model adds A.

Table 1: Evidence supporting the reward-selection analysis. Cohorts can share trained policies; group-level reweighting holds recorded samples fixed.

Study	Coverage	Role in the argument
Reward groups	2,812 groups, 74 runs	Distinguish answer availability from selection
Reward intervention	288 runs, 5 weights	Measure the response to oracle anchoring
Prefix diagnostics	74 self, 131 mixture runs	Test information beyond response accuracy
Model comparison	30 runs, 4 models	Describe policy-specific audit–gain relations

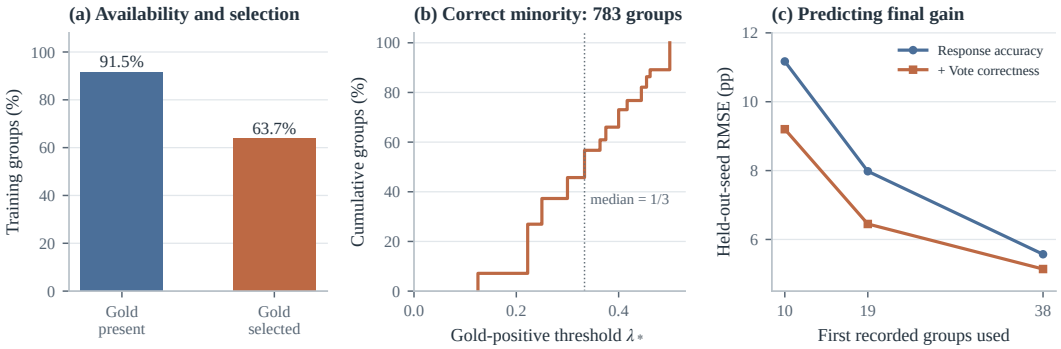

Figure 2: From available answers to supervision. (a) Correct-answer availability and correct plurality selection in 2,812 groups from 74 self-rewarded runs. (b) Distribution of the weight λ_* above which correct responses obtain positive centered reward in the 783 groups with a wrong plurality and an available correct answer. (c) Leave-one-training-seed-out prediction of final pass@1 gain using fixed prefixes of the recorded groups. Both models use response accuracy; the augmented model adds vote correctness.

Mixture comparisons include categorical reward-weight indicators. These retrospective analyses use the same screened split and final evaluation outcomes; the held-out units are training seeds. Prefix sizes and model forms are fixed across folds, with all evaluated windows reported.

Matched information budgets. The baseline and augmented predictors use the same runs, responses, gold answers and observation window. Adding A changes how the observed answers are summarized: P counts correct responses, whereas A records whether correctness reaches the selected supervision target. Neither predictor uses the remaining groups when a prefix is evaluated. The target is final held-out gain relative to the frozen policy. This design isolates the information carried by selection from access to more samples or more gold labels, while the reward-mixture experiment separately changes the supervision used for learning.

5 Experimental Results

5.1 Available Answers and Their Reinforcement Signals

Across 2,812 recorded majority-reward groups, 2,574 contain a correct answer and 1,791 select it. The difference is 783 groups, or 27.8% of all groups, where a correct response is available but receives negative centered majority reward. Of the remaining groups, 238 contain no correct response and 74 are unanimously correct; the latter have zero centered reward. This decomposition locates the selection effect among groups containing usable correct responses.

Applying Proposition 1 to the same 783 groups gives thresholds from $1/8$ to $1/2$, with median $1/3$ (fig. 2b). At $\lambda = 0.25$, 211 of these groups assign positive centered reward to the correct response, 81 assign zero, and 491 retain negative reward. At $\lambda = 0.50$, 698 assign positive reward and 85 assign zero. At $\lambda = 0.75$, all 783 assign positive reward. Table 2 summarizes this progression. These are exact counterfactual rewards on fixed recorded groups. They identify how anchoring changes available supervision before considering the response of a trained policy.

Table 2: Centered reward of correct responses in the same 783 wrong-vote groups containing a gold answer. Reweighting changes supervision on fixed samples; training outcomes are measured separately in Table 3.

Oracle weight	0	0.25	0.50	0.75
Positive	0	211	698	783
Zero	0	81	85	0
Negative	783	491	0	0

Table 3: Reward intervention on the fixed Llama-3-8B split. Gains are changes in held-out pass@1; standard deviations are across runs.

Reward	Runs	Mean gain	s.d.	Gain range
Majority ($\lambda = 0$)	74	$+0.031$	0.131	$[-0.23, +0.24]$
Mixture ($\lambda = 0.25$)	49	$+0.057$	0.118	$[-0.21, +0.23]$
Mixture ($\lambda = 0.50$)	39	$+0.146$	0.053	$[-0.01, +0.23]$
Mixture ($\lambda = 0.75$)	43	$+0.184$	0.041	$[+0.05, +0.24]$
Oracle ($\lambda = 1$)	83	$+0.196$	0.031	$[+0.12, +0.26]$

Table 4: Predicting final pass@1 gain from the first 19 of 38 recorded groups. Leave-one-seed-out RMSE is in percentage points. Each augmented model adds selection correctness A to the listed controls; mixture models also control reward weight.

Cohort	Controls	Without A	With A	Reduction
Majority	Response accuracy P	7.98	6.45	19.2%
	P and P^2	7.69	6.50	15.4%
	P and five signals	7.37	6.78	8.0%
Mixture	Response accuracy P	7.14	5.99	16.1%
	P and P^2	7.11	6.03	15.2%
	P and five signals	6.30	6.01	4.6%

5.2 Reward Anchoring and Adaptation Gains

Oracle anchoring increases the mean gain and concentrates outcomes. Table 3 reports 288 runs under the shared configuration. Every oracle-rewarded run improves, and the ratio of majority to oracle standard deviation is 4.19. The intermediate weights connect the two endpoints: mean gain increases at each tested weight and standard deviation decreases. The groupwise threshold in Proposition 1 identifies the change in centered reward that accompanies this intervention.

Partial anchoring recovers much of the oracle response. At $\lambda = 0.50$, the mean gain is 0.146 and the standard deviation is 0.053. At $\lambda = 0.75$, the mean reaches 0.184, or approximately 94% of the oracle endpoint's mean gain, while the standard deviation is 69% smaller than in the majority arm. The intermediate weights trace a gradual increase in gain and decrease in variability between the two endpoints. Here λ is a reward weight, not a fraction of labeled training examples: all mixture conditions use gold answers to construct their weighted reward.

5.3 Selection Correctness Predicts Final Gain from Training Prefixes

Information beyond response accuracy. Using the first 19 recorded groups, response accuracy alone predicts final pass@1 gain with RMSE 7.98 percentage points in the majority cohort. Adding selection correctness reduces it to 6.45, a 19.2% reduction (table 4). The corresponding reduction in the mixture cohort is 16.1%, from 7.14 to 5.99 points, after controlling reward weight. Both models observe the same sampled responses and gold answers. The additional information comes from identifying which answer wins the vote.

The selection signal persists across accuracy controls. A quadratic response-accuracy baseline tests whether A merely adds curvature to the relationship between P and gain. The error reductions remain 15.4% and 15.2% in the majority and mixture cohorts. Adding the five concentration signals yields reductions of 8.0% and 4.6%. Across 20 repeated five-fold seed-grouped partitions, adding A improves pass@1 prediction in every partition for both the accuracy-only and expanded controls. These repetitions test sensitivity to fold assignment on the same runs. Table 5 quantifies the corresponding uncertainty.

Observation budget changes the incremental value. With the first ten groups, adding A to P reduces error by 17.6% in majority runs and 9.3% in mixtures. With all 38 groups, the reductions are 7.7% and 11.1%. For the full-trajectory baseline combining P and all five signals, the corresponding reductions are 0.7% and 2.0%. The prefix comparisons therefore locate the diagnostic's most useful regime: selection correctness adds information while the recorded trajectory is still incomplete. Tables 18 and 5 report all windows and both outcome budgets.

Uncertainty and fold sensitivity. The gain from selection correctness can be examined at the level of prediction errors, rather than through correlations alone. For each leave-one-seed-out fit, we retain the paired baseline and augmented predictions and resample training-seed clusters. With 19 groups, the descriptive 95% intervals for RMSE improvement over response accuracy are $[0.54, 2.52]$ points for majority and $[0.51, 1.72]$ for mixtures (table 5). The quadratic accuracy comparisons likewise retain positive intervals. The expanded accuracy-plus-concentration controls yield smaller point improvements, with intervals spanning zero. These comparisons distinguish the stronger contribution beyond response accuracy from the finer increment after all observed concentration features are supplied.

Table 5: Descriptive 95% seed-cluster intervals for the RMSE improvement from adding A, using 19 groups. Predictions are held fixed during bootstrap.

Cohort	Controls	pass@1	pass@4
Majority	P	$[0.54, 2.52]$	$[0.42, 2.46]$
Majority	P, P^2	$[0.29, 2.09]$	$[0.03, 1.88]$
Majority	P + five signals	$[-0.06, 1.24]$	$[-0.12, 1.17]$
Mixture	P	$[0.51, 1.72]$	$[0.38, 1.51]$
Mixture	P, P^2	$[0.47, 1.64]$	$[0.28, 1.29]$
Mixture	P + five signals	$[-0.02, 0.58]$	$[-0.10, 0.37]$

The repeated five-fold analysis changes how runs are partitioned while keeping features and observation budgets fixed. Across the 20 partitions, mean RMSE for the response-accuracy comparison changes from 8.00 to 6.47 points in majority runs and from 7.14 to 6.01 in mixtures. With expanded controls, the means change from 7.54 to 6.90 and from 6.32 to 6.05 points. The direction is consistent across partitions, providing a check that the observed increment is not specific to leaving out one seed at a time. All of these comparisons condition on the recorded runs and shared problem split. The intervals resample fixed out-of-fold errors rather than refitting models; the exact procedure appears in section H.

The diagnostic persists under numeric-equivalent scoring. We additionally score stored parsed decimal answers by exact rational value, so, for example, `12.00` and `12` agree. This changes endpoint scoring while retaining the rewards actually used during training. The majority mean gain is 0.028 with s.d. 0.134; the oracle mean remains 0.196 with s.d. 0.031, and all 83 oracle runs improve. With these rescored pass@1 targets, the 19-group response-accuracy comparison changes RMSE from 8.20 to 6.63 points for majority and 7.29 to 6.11 for mixtures. This check operates on stored parsed answers; it does not reparse the original generated text or retrain with a different reward parser.

The audit resolves much of the variability within self-reward. The majority arm has Pearson correlation $r = 0.918$ between A and held-out pass@1 gain (fig. 5a). Its linear fit is $\widehat{\Delta} = 0.729A - 0.434$, with a zero crossing at 0.595. The bin means increase from -0.195 below $A = 0.35$ to $+0.166$ above 0.75 (table 9). The fitted crossing summarizes this model and protocol; its diagnostic variable is measured during training.

Table 6: Leave-one-training-seed-out prediction of held-out accuracy gains in 74 majority-rewarded runs. RMSE is in percentage points; lower is better. The five-signal block contains margin, plurality share, distinct-answer count, entropy and the identical-answer-group fraction.

Target	Signals	Without audit	With audit	Reduction
pass@1	Margin	10.02	5.32	46.9%
pass@1	Five-signal block	9.84	5.48	44.3%
pass@4	Margin	11.31	6.13	45.8%
pass@4	Five-signal block	11.43	6.44	43.6%

Table 7: Model-specific majority-reward cohorts. Each model has its own screened 200/100 split. Audit and gain are means over the listed runs; r is the within-model audit–gain correlation.

Model	Runs	Base pass@1	Audit	Gain	r
Llama-3-8B-Instruct	12	0.43	0.612	+0.005	0.897
Gemma-2-2B-it	6	0.41	0.715	+0.207	0.793
Qwen3-8B	6	0.65	0.776	+0.012	0.475
Qwen3-0.6B	6	0.22	0.294	+0.063	0.910

Vote correctness is more closely associated with gain than margin ($r = 0.656$), plurality share (0.666), entropy (-0.621), or distinct-answer count (-0.582; table 13). Controlling for mean margin, the partial correlation remains 0.85, and adding audit to the margin regression increases R^2 by 0.42 (run-bootstrap 95% interval $[0.24, 0.61]$). Correctness thus captures information about the supervision beyond how concentrated the votes are. The regressions describe associations along training trajectories; the λ comparison directly intervenes on the reward.

Audit information transfers to held-out runs. We test the diagnostic on runs excluded from regression fitting, using leave-one-training-seed-out ordinary least squares. Each model uses either margin or the five-signal block, with and without audit. The model form and signals remain fixed across folds. Table 6 reports the resulting errors on all 74 majority-rewarded runs. Adding audit to the five-signal block reduces pass@1 RMSE from 9.84 to 5.48 percentage points, a 44.3% reduction. For pass@4, RMSE falls from 11.43 to 6.44 points, a 43.6% reduction. Thus correctness adds information about unseen runs beyond the combined vote-concentration signals. These predictions use training-time diagnostics on the fixed screened split. The corresponding in-sample fits appear in table 16.

The relationship survives influence and reward-weight controls. Removing any one majority run leaves the audit–gain correlation between 0.914 and 0.927. Ordinary least squares, Huber and Theil–Sen fits give slopes of 0.729, 0.734 and 0.736, respectively. The association therefore persists under different treatments of influential observations. In the 131 mixture runs, controlling for reward weight and the five-signal block, audit reduces leave-one-seed-out pass@1 RMSE from 7.06 to 4.73 points (33.1%). This comparison groups shared seeds within each fold and separates audit information from differences in mean gain across reward weights. Appendix F.1 gives the protocol and both evaluation budgets.

5.4 Model Dependence

Four policy-specific response curves. The cross-model cohorts in table 7 and fig. 3 contain 30 runs from Llama, Gemma and Qwen. Audit and gain are positively associated within each cohort, with correlations from 0.475 to 0.910. The fitted response is model dependent: slopes range from 0.206 to 0.793 units of gain per unit audit. The Qwen3-0.6B fit crosses zero near 0.21, while the 74-run Llama fit crosses near 0.60. Each fit relates the audit to gain for one model on its own screened problem pool.

A joint view of policy quality and reward quality. The Llama and Gemma cohorts have similar base pass@1 values, 0.43 and 0.41, while their mean audits are 0.612 and 0.715. Their mean gains are

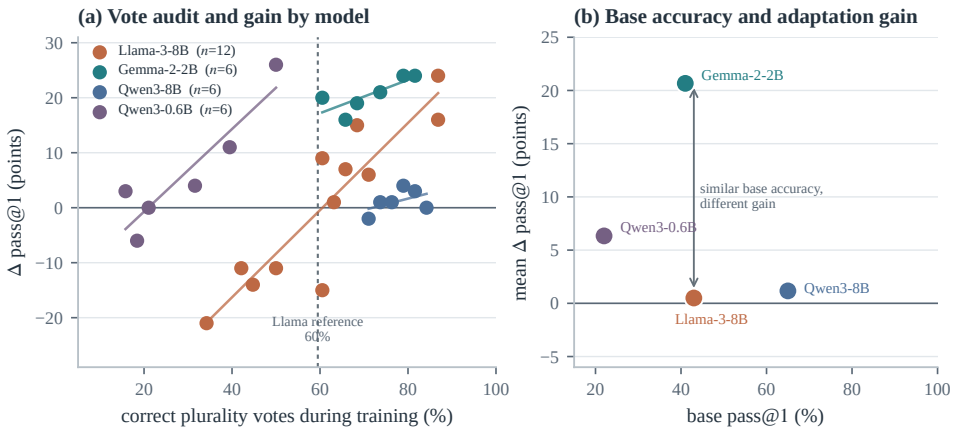

Figure 3: (a) Audit–gain associations within four policies from three model families, with a separate descriptive fit for each. (b) Mean gain against base accuracy. Similar base accuracy can accompany different adaptation gains; the model-specific fits describe different reward trajectories.

0.005 and 0.207, respectively. This contrast illustrates why starting accuracy alone is an incomplete description of an adaptation experiment. The comparison becomes more informative when base accuracy is considered alongside the correctness of the supervision encountered during training. Because each model has its own screened pool, the comparison reflects differences in both the policy and the selected problems.

6 Discussion

Answer supply and reinforcement targets. The group decomposition separates a model's ability to generate a correct answer from the reward rule's selection of that answer. The $H - A$ gap identifies groups where correct responses are available yet receive adverse centered majority rewards. The threshold λ_g^* then quantifies how much oracle anchoring changes that sign for each observed group. Together these measurements describe the supervision that lies between sampled answers and parameter updates.

Diagnostic use during adaptation. A prefix audit answers whether the correct responses observed so far are also winning the votes used for supervision. Its incremental value over response accuracy supports monitoring reward selection alongside policy performance. The current analysis uses gold labels on screened direct-answer GSM8K and holds out training seeds on a fixed problem split. Cross-model cohorts use model-specific screened pools. These settings define the measured diagnostic relationships; implementing a gold-free controller would require an observable estimate of selection correctness.

Supplementary studies report finite-budget sampling curves, paired input edits and margin-gate admission counts (section G). They provide additional views of the reward channel, with their individual protocols and metrics stated alongside the results.

7 Conclusion

Correct answers become supervision through a selection step. We quantify the gap between answer availability and correct plurality selection, derive the oracle weight that changes the centered reward of a correct minority response, and measure the training response across five reward weights. Prefix audits add information about final adaptation gains beyond response accuracy, including under nonlinear accuracy controls. The resulting account connects what a policy can sample, what its reward rule reinforces, and how its adaptation unfolds.

AI USE DISCLOSURE

AI assistants supported writing, literature search, implementation, analysis and derivation checking. The authors are responsible for the manuscript and its reported evidence.

REPRODUCIBILITY STATEMENT

The appendix specifies the two training instruments, reward callback audit, metric estimators, sampling settings and result cohorts. CPU analysis scripts recompute aggregates from recorded run outputs and regenerate the quantitative figures. Model-based training uses TRL GRPO for the GSM8K study and an explicit streaming policy-gradient loop for the admission experiment.

ETHICS STATEMENT

Experiments use locally hosted models and public or procedurally generated mathematics problems. No deployed service or private user data is targeted. The stress tests assume write access to training prompts. Their edit budget measures surface similarity, and empirical gates are evaluated through observable admission and held-out metrics.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv:2107.07511*, 2021.

Karl Cobbe et al. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.

Jeremy Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. In *arXiv:1708.06733*, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Vanshaj Khattar, Md Rafi ur Rashid, Moumita Choudhury, Jing Liu, Toshiaki Koike-Akino, Ming Jin, and Ye Wang. Amplification effects in test-time reinforcement learning: Safety and reasoning vulnerabilities. *arXiv preprint arXiv:2603.15417*, 2026.

Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Ruotong Liao, Nikolai Röhrich, Xiaohan Wang, Yuhui Zhang, Yasaman Samadzadeh, Volker Tresp, and Serena Yeung-Levy. Tool verification for test-time reinforcement learning. *arXiv preprint arXiv:2603.02203*, 2026.

Hongxiang Lin, Zhirui Kuai, Erpeng Xue, and Lei Wang. Detecting and mitigating the correct-answer extinction window in test-time reinforcement learning with majority voting. *arXiv preprint arXiv:2605.19444*, 2026.

Pascal Massart. The tight constant in the dvoretzky–kiefer–wolfowitz inequality. *The Annals of Probability*, 1990.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Javier Rando and Florian Tramèr. Universal jailbreak backdoors from poisoned human feedback. In *International Conference on Learning Representations (ICLR)*, 2024.

Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Zhihong Shao et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. In *arXiv:2402.03300*, 2024.

David Silver and Richard S. Sutton. Welcome to the era of experience. Manuscript, 2025.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Weiqin Wang, Yile Wang, Kehao Chen, and Hui Huang. Beyond majority voting: Towards fine-grained and more reliable reward signal for test-time reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

Dong Yan, Jian Liang, Yanbo Wang, Shuo Lu, Ran He, and Tieniu Tan. What if consensus lies? selective-complementary reinforcement learning at test time. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

Yongcan Yu, Lingxiao He, Jian Liang, Kuangpu Guo, Meng Wang, Qianlong Xie, Xingxing Wang, and Ran He. Understanding and mitigating spurious signal amplification in test-time reinforcement learning for math reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *International Conference on Machine Learning (ICML)*, 2024.

Yang Yue et al. Does reinforcement learning really incentivize reasoning capacity beyond the base model? *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Todd Y. Zhou. When RLVR shrinks the reasoning boundary: Diagnosing Pass@k inversion. *arXiv preprint arXiv:2607.20543*, 2026.

Yuxin Zuo et al. Ttrl: Test-time reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

A Related Work

Learning from self-generated rewards. TTRL constructs pseudo-rewards from majority votes (Zuo et al., 2025), connecting to self-rewarding models (Yuan et al., 2024) and experience learning (Silver & Sutton, 2025). Confidence weighting (Wang et al., 2026), denoising ambiguous signals (Yu et al., 2026), selective/complementary pseudo-labeling (Yan et al., 2026) and tool verification (Liao et al., 2026) intervene on the quality of this supervision. Our controlled mixture quantifies the reward response while holding the optimization protocol fixed.

Training dynamics and sampling efficiency. Lin et al. (2026) analyze correct-answer extinction, label migration and online safeguards in majority-rewarded TTRL. Our training-time audit shares the interest in vote correctness; the contribution here is the controlled oracle-mixture response and its connection to paired input perturbations. Yue et al. (2025) analyze the difference between small- and large-budget sampling under RLVR, with pass@k inversion also studied by Zhou (2026). Our coverage curves describe this distinction on the screened self-rewarded population. Noisy-label learning (Natarajan et al., 2013) supplies a related perspective on unreliable targets.

Poisoning and selective updates. Poisoning and backdoors are studied in supervised learning (Gu et al., 2017; Shafahi et al., 2018; Li et al., 2022) and human-feedback training (Rando & Tramèr, 2024). In TTRL, harmful prompt injections can amplify behaviors and affect reasoning (Khattar et al., 2026). Our intervention uses bounded edits of the training questions and relates the resulting gain differences to training-time vote correctness. Formal robustness frameworks include randomized smoothing (Cohen et al., 2019), conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021), and concentration bounds (Massart, 1990). The threshold-saturation observation here is specific to the empirical calibration rule in (8).

B Analytical Results

Remark 1 (Fixed-target crossover)**.** *When $\hat{y}_N \neq y^*$ and both answers are available, the two answers receive rewards $1-\lambda$ and λ. Their ordering switches at $\lambda = 1/2$, where they tie. The largest available reward is $\max\{\lambda, 1-\lambda\}$; when the vote is correct, it is 1.*

A KL-regularized distributional problem makes the local effect explicit. For fixed targets and a fixed penalty $\eta > 0$, maximize $\mathbb{E}_q[r_\lambda] - \eta\,\mathrm{KL}(q\|q_0)$ over answers in the base support. Its solution is

$$q_\lambda(a) = \frac{q_0(a)\exp(r_\lambda(a)/\eta)}{Z_\lambda}. \tag{6}$$

Proposition 2 (Local reward sensitivity)**.** *For a wrong plurality label with positive base probability on it and on the gold answer, $q_\lambda(y^*)$ increases and $q_\lambda(\hat{y}_N)$ decreases with λ at fixed η. For any q satisfying $\mathrm{KL}(q\|q_0) \leq \rho$, its expected mixture reward obeys*

$$R_\lambda(q) \leq R_\lambda(q_0) + \sqrt{\rho/2}, \qquad R_\lambda(q) = \mathbb{E}_q[r_\lambda]. \tag{7}$$

These results hold for fixed targets (proofs in section B). GRPO refreshes votes and shares parameters across problems, so we measure its response experimentally. The bound applies to expected reward; the sampling curves separately measure finite-budget answer coverage.

Fixed-target reward envelope. For a correct plurality label, the mixture assigns reward one to that answer. For an incorrect label with both answers in the support, it assigns λ to the gold answer and $1-\lambda$ to the plurality answer. Hence the available reward envelope is 1 or $\max(\lambda, 1-\lambda)$, respectively. If a fixed-label item population has vote correctness m, its average envelope is $m + (1-m)\max(\lambda, 1-\lambda)$. This is a reward envelope, not an accuracy ceiling.

Proof of Proposition 2. Stationarity of the normalized KL-penalized objective gives (6). For a wrong-vote item define $h(a) = \mathbf{1}[a = y^*] - \mathbf{1}[a = \hat{y}_N]$. Differentiating at fixed η yields

$$\frac{\partial q_\lambda(a)}{\partial \lambda} = \frac{q_\lambda(a)}{\eta}\big(h(a) - \mathbb{E}_{q_\lambda} h\big).$$

Table 8: GSM8K training and evaluation configuration. The synthetic streaming instrument is specified separately in section E.

Setting	Value
Main model	Llama-3-8B-Instruct
Additional models	Gemma-2-2B-it, Qwen3-8B, Qwen3-0.6B
Training / evaluation items	200 / 100, screened separately per model
Screening / training group sizes	4 / 8
Endpoint / coverage samples per item	4 / 32
Optimization steps	300
Learning rate / warmup	3×10^{-6} / 0.05
KL coefficient / gradient accumulation	0.02 / 1
LoRA rank / alpha / dropout	16 / 32 / 0.05
LoRA modules	query and value projections
Quantization	4-bit NF4, double quantization
Endpoint evaluation item batch	2
Split / evaluation seed	20260910 / 20260911

The derivative at the gold answer is positive and at the wrong plurality answer negative when both have positive probability, since $-1 < \mathbb{E}h < 1$. A hard KL radius has a multiplier that can vary with λ; the derivative statement holds for the fixed penalty specified above. For the second claim, $0 \le r_\lambda \le 1$ implies $|R_\lambda(q) - R_\lambda(q_0)| \le \mathrm{TV}(q, q_0)$. Pinsker's inequality gives (7). □

Sampling response. For independent responses with per-sample correctness p, $\mathrm{pass@}k = 1 - (1 - p)^k$ and its derivative is $k(1-p)^{k-1}$. Thus small changes in p can have different effects depending on k and saturation. For a fixed distribution, the limit as $k \to \infty$ records whether $p > 0$; a measurement at $k = 32$ remains a finite-budget statistic.

Proof of Proposition 3. Since $d \le 1$, the strict inequality $d > \tau$ is impossible if $\tau \ge 1$. The second term in (8) is at least $2\widehat{L}\varepsilon$, which proves the sufficient saturation condition. □

C GSM8K Protocol and Recorded Cohorts

Prompt and sampling. Screening, GRPO training and endpoint evaluation use the same direct-answer instruction, rendered through each model's chat template:

> Solve the arithmetic problem. Give only the final numeric answer inside `\boxed{}`; do not include intermediate answers.

Evaluation uses temperature 0.8 and a 256-token cap. The GRPO invocation uses the trainer's generation defaults apart from invalid-value removal and logit renormalization; it does not explicitly set the evaluation temperature for training. Qwen thinking is disabled in prompt rendering. The sampled response is parsed by extracting a boxed answer or the last numeric match after light LaTeX normalization. Rewards and the primary metrics use exact equality of these parsed strings. Section 5.3 additionally checks decimal-equivalent endpoint scoring on the stored parsed answers. Plurality ties in the recorded runs use the runtime ordering of `max(set(valid), key=valid.count)`; the chosen label is retained in the audit. The example in table 12 contains a tie and reports that recorded choice.

Metric estimator. For M problems and binary sample outcomes z_{ij}, the reported estimator is $\widehat{P}_k = M^{-1} \sum_i \mathbf{1}[\sum_{j=1}^{k} z_{ij} > 0]$. It uses a fixed prefix of independent sampled responses rather than the all-subsets combinatorial estimator. Its increments on the 100-item set have granularity 0.01. Endpoint evaluation batches two items with four responses each; the separate 32-response coverage evaluation batches one item. The frozen pass@1 values 0.43 and 0.34 therefore refer to different sampled response sets. Within each evaluation procedure, pre/post seeds and batching match.

Audited groups. `reward_audit` is updated inside the GRPO reward function called during `trainer.train()`; evaluation does not update it. Each retained majority run has 38 callback groups, with rows containing the gold, plurality, vote counts and correctness. The generation batch contains eight responses, while optimizer minibatches contain one response with gradient accumulation one. A generation group thus supplies eight minibatches; 300 steps require $\lceil 300/8 \rceil = 38$ groups, with the final group used partially. Audit rows follow callback append order and weight each generated group equally. The saved group index is local to the callback, so prefix analyses use row order, not that index. The records retain gold answers and counts, rather than problem IDs or per-group timestamps. The reported zero-std diagnostic is the fraction with identical parsed answers, as counted by the callback, rather than the variance of mixture rewards.

Cohort selection. The endpoint and mixture cohort contains the 288 distinct arm/seed/weight records matching the model, screening summary, split/evaluation seeds, item counts, group size, step count, cap, prompt mode, learning rate, KL coefficient, evaluation batching, QLoRA and binary reward configuration in table 8. The cross-model cohort uses the 12 Llama replication runs and six runs for each other model. The coverage cohort contains one frozen and 56 adapter evaluations on Llama. The 23 connection cells use the same screened Llama split and paired clean/perturbed arms. A CPU analysis script enumerates the exact source files and recomputes the reported cohort statistics.

Statistical summaries. Run-level means, sample standard deviations and correlations describe the fixed screened splits. Bin-mean intervals in table 9 are normal approximations over runs. The incremental-regression bootstrap resamples runs. These summaries condition on the common evaluation set rather than treating repeated seeds as new test problems. The fits in table 7 describe the audit–gain relationship within each model's screened pool.

D Input Perturbations

Candidates are enumerated in this order: percentage restatement, unit conversion, alias substitution, lexical replacement. The default similarity is Jaccard overlap of lowercased whitespace-token sets; an embedding option exists but is not enabled by default. The first admissible candidate with similarity at least $1 - \varepsilon$ is used. No search over victim rewards or gold labels chooses the candidate.

Percentage transformations include "25% more than" to "1.25 times as many as"; unit transformations include "2 hours" to "120 minutes". Aliases replace a supplied occurrence, and lexical candidates include "Calculate" to "Determine", "produced" to "manufactured", and "sold" to "offered". The last example can change a word problem's meaning, so the library is a surface-perturbation generator, not a semantic verifier. The recorded GSM8K connection experiment retains the original gold annotation without recomputing it. Its 109 modified prompts comprise 84 lexical and 25 unit-conversion cases. Interpretation of these cells concerns sensitivity to this concrete input-edit procedure.

E Streaming Instrument and Calibration

Stationary streams. The admission study uses procedurally generated addition questions with four numeric choices. It processes 32 questions, with the first half selected for perturbation. The gate and no-gate branches start from the same model and seed; their trajectories can diverge after accepting different updates. The code selects the first admissible edit, and the addition prompt admits a "Calculate" to "Determine" substitution.

The policy-gradient loop uses four responses per question, temperature 0.8, a 32-token cap, learning rate 10^{-5}, reference coefficient 0.003, and LoRA rank 4/alpha 8 on query and value projections. It centers binary plurality rewards within the group and sums response-token log probabilities. Its reference term is the sampled log-probability difference from the initial policy. This explicit streaming instrument differs from the TRL GRPO loop used for the GSM8K reward study. Clean evaluation averages the correctness flags across sampled held-out responses; it is sample accuracy, not a majority-vote score.

Admission counts and uncertainty. For each perturbed item, the current policy samples the clean prompt to define eligibility C_i, then samples the edited prompt to obtain the training label. The

result's `label` is from that training rollout. Conditional counts in table 17 consequently measure admission of wrong supervision. Gate and no-gate denominators can differ because their current policies differ. The Wilson intervals in fig. 6 summarize pooled attempts; multiple attempts share a trained policy, so they are descriptive rather than a cluster-adjusted significance test. We report the direction of change for each matched cell and measure held-out accuracy separately to assess learning outcomes.

Calibration rule. For m clean calibration margins, the implementation sets $u = \max\{\alpha - \sqrt{\ln(2/\alpha)/(2m)}, 0\}$ and uses order statistic $\widehat{Q}_\alpha^{\text{low}} = d_{(\max\{1,\lceil mu\rceil\})}$. This specifies the historical calibration statistic, including its clipping at the sample minimum; it is not a guaranteed population lower quantile when $u = 0$. The paired sensitivity statistic is

$$\widehat{L}_{\text{raw}} = \max_i \frac{1}{2N\varepsilon} \sum_a |c_i(a) - c_i'(a)|.$$

The selected $\widehat{L}$ is the smallest grid value in $\{0.5, 1, 2, 4\}$ at least as large as this statistic, or the statistic itself when it exceeds four. This yields (8), using $\sqrt{\ln(1/\alpha)/(2N)}$ as a separate sampling-scale term. Both the distance normalization by the allowed budget and the finite observed maximum make $\widehat{L}$ an empirical proxy, not a global Lipschitz bound. The recorded 128-item calibration has zero clean quantile at the reported N values, so its thresholds are determined by the second term.

E.1 Finite-audit threshold calibration

The calibration implementation also evaluates a threshold with an empirical perturbation-sensitivity term:

$$\tau_{\text{cal}} = \max\left\{\widehat{Q}_\alpha^{\text{low}},\, 2\widehat{L}\varepsilon + \sqrt{\frac{\ln(1/\alpha)}{2N}}\right\}. \tag{8}$$

Here $\widehat{Q}_\alpha^{\text{low}}$ is a clean-margin order statistic, and $\widehat{L}$ is selected to cover the largest observed paired answer-frequency change divided by the edit budget. The quantile adjustment and the separate N-dependent term are defined in section E. The finite paired audit supplies an empirical sensitivity estimate.

Proposition 3 (Threshold saturation)**.** *A gate using $d > \tau$ with $d \in [0, 1]$ accepts no update when $\tau \geq 1$. For* (8)*, $2\widehat{L}\varepsilon \geq 1$ is sufficient for this saturation regardless of N.*

On a Qwen2.5-1.5B addition-problem calibration set of 128 items, $\varepsilon = 0.20$ and $\alpha = 0.05$, the N-dependent term decreases from 0.433 to 0.153 between $N = 8$ and 64. The resulting thresholds are $2.43, 1.91, 1.84, 1.75$ at $N = 8, 16, 32, 64$ (fig. 4). All four thresholds exceed the largest vote margin and therefore reject every update. The fixed thresholds evaluated in the streaming study provide operating points that admit updates. A distribution-free guarantee would require a bound over the full allowed perturbation set and a specified safety event; the finite paired audit estimates empirical sensitivity.

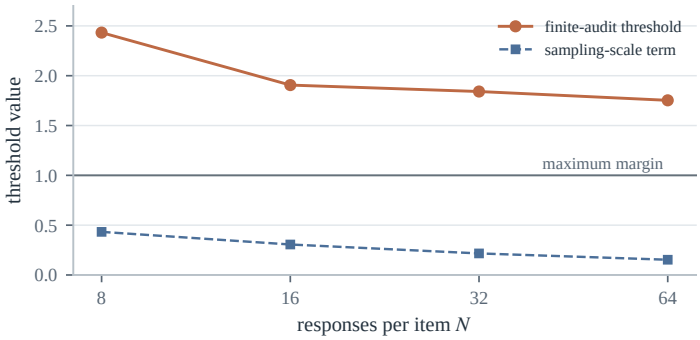

Figure 4: Finite-audit calibration on Qwen2.5-1.5B. The implemented threshold (8) remains above one as the sampling term contracts. The empirical sensitivity term keeps the threshold above the maximum vote margin, so the calibrated gate admits no updates at these settings.

F ADDITIONAL RESULTS

The following tables give audit-stratified outcomes, sampling curves, and comparisons between diagnostic signals.

Table 9: Endpoint study binned by training-time vote audit. Each delta is evaluated on the same 100-item split, with granularity 0.01. Intervals summarize bin means over runs.

Arm	vote audit	seeds	mean Δpass@1	s.d.	95% CI
majority	0.00–0.35	6	-0.195	0.031	$[-.220, -.170]$
majority	0.35–0.45	6	-0.137	0.015	$[-.149, -.125]$
majority	0.45–0.55	9	-0.101	0.041	$[-.128, -.074]$
majority	0.55–0.65	17	$+0.028$	0.073	$[-.007, +.062]$
majority	0.65–0.75	18	$+0.095$	0.055	$[+.070, +.120]$
majority	0.75–1.00	18	$+0.166$	0.058	$[+.139, +.192]$
majority	all	74	$+0.031$	0.131	$[+.001, +.060]$
oracle	—	83	$+0.196$	0.031	$[+.189, +.203]$

Table 10: Representative coverage evaluations using 32 responses per problem. Reported pass@k uses the first k samples.

Run	pass@1	pass@2	pass@4	pass@8	pass@16	pass@32
base policy	0.34	0.53	0.69	0.85	0.93	0.97
majority reward	0.54	0.71	0.89	0.95	0.97	0.97
oracle reward	0.52	0.69	0.85	0.93	0.98	0.98
Δ majority	$+0.20$	$+0.18$	$+0.20$	$+0.10$	$+0.04$	0.00
Δ oracle	$+0.18$	$+0.16$	$+0.16$	$+0.08$	$+0.05$	$+0.01$

Table 11: Coverage deltas for the 32 policies with pass@1 gain above two percentage points. Values are percentage points relative to the frozen policy, with medians and observed ranges.

k	median Δpass@k	range
1	$+8.0$	$+3$ to $+20$
2	$+6.5$	$+2$ to $+18$
4	$+5.0$	$+1$ to $+20$
8	$+4.0$	$+2$ to $+10$
16	$+2.0$	0 to $+5$
32	0.0	0 to $+1$

Diagnostics across the reward-mixture cohort. Table 14 extends the signal comparison to the 131 runs at $\lambda \in \{0.25, 0.50, 0.75\}$. Across these pooled reward weights, the audit has Pearson correlation 0.849 with pass@1 gain. Controlling for margin yields partial correlation 0.721; adding audit raises the margin regression's R^2 from 0.429 to 0.726, and the five-signal regression's R^2 from 0.463 to 0.730. These summaries describe joint variation across weights and trajectories. The fixed-majority results in the main text isolate the diagnostic comparison within one reward construction.

F.1 CROSS-RUN VALIDATION AND INFLUENCE ANALYSIS

The diagnostic regressions use an intercept and ordinary least squares, with no hyperparameter fitting or feature selection. The reduced models use either margin or the five-signal block; the augmented model adds audit. For each training seed, we fit on all other seeds and predict every record with that seed. The majority cohort contains 74 distinct seeds. The mixture cohort contains 131 records from 119 seeds, and all records sharing a seed are held out together. Both mixture models include categorical indicators for reward weights 0.50 and 0.75, using 0.25 as the reference. Errors are pooled

Table 12: Recorded vote-audit example ($N = 8$). Answers 20 and 600 tie at two votes each; the stored plurality is 20. Both differ from gold 1200, so this group has audit correctness zero.

answer	0	120	1200	20	360	600
votes	1	1	1	2	1	2

Table 13: Training-time diagnostic signals over 74 majority-rewarded runs. Pearson correlations are shown for pass@1 and pass@4 gain, and Spearman correlation for pass@1 gain.

Training signal	r (Δpass@1)	ρ (Δpass@1)	r (Δpass@4)
vote audit	$+0.918$	$+0.904$	$+0.901$
mean vote margin	$+0.656$	$+0.676$	$+0.581$
plurality share	$+0.666$	$+0.688$	$+0.600$
mean distinct answers	-0.582	-0.604	-0.543
mean group entropy	-0.621	-0.648	-0.567
zero-std fraction	$+0.133$	$+0.139$	$+0.144$

Table 14: Diagnostic signals in the 131 reward-mixture runs, pooled across three oracle weights. Correlations use held-out pass@1 gain.

Training signal	Pearson r	Spearman ρ
Vote audit	0.849	0.763
Mean vote margin	0.655	0.720
Plurality share	0.627	0.691
Mean distinct answers	-0.473	-0.535
Mean group entropy	-0.540	-0.611
Identical-answer-group fraction	0.210	0.265

over left-out records, with one prediction per record. All runs retain the original screened problem split, so this evaluates cross-run diagnostic performance.

Table 15 reports the weight-controlled comparisons. Audit reduces error for both signal blocks and both evaluation budgets. Within the individual mixture weights 0.25, 0.50 and 0.75, audit–gain Pearson correlations are 0.938, 0.573 and 0.676, respectively, over 49, 39 and 43 runs. The association is therefore present within each tested weight as well as in the pooled cohort.

For the majority cohort, deleting each run in turn gives a Pearson correlation range of $[0.914, 0.927]$. The audit coefficient in the full five-signal regression remains positive, between 0.660 and 0.712, across all deletions. The univariate Huber fit uses a transition scale of 1.4826 times the median absolute deviation of the initial OLS residuals (0.04965 in accuracy units). Its slope is 0.734, compared with OLS 0.729 and Theil–Sen 0.736. These checks retain all observations in the reported main fits and show that no single run determines the positive association. Audit is measured during training; the validation evaluates its diagnostic information across runs.

In-sample regression comparison. The nested fits in table 16 complement the cross-run predictions. In the margin-plus-audit pass@1 fit, the audit coefficient is 0.693 per unit audit, with a run-bootstrap 95% interval $[0.590, 0.788]$. A ten-point audit difference corresponds to about 6.9 points of fitted gain at fixed margin. For pass@4, audit–gain Pearson correlation is 0.901 and the partial correlation controlling for margin is 0.847.

Table 15: Leave-one-training-seed-out errors in 131 mixture runs. Both models include categorical reward-weight controls. RMSE is in percentage points.

Target	Signals	Without audit	With audit	Reduction
pass@1	Margin	7.00	4.58	34.6%
pass@1	Five-signal block	7.06	4.73	33.1%
pass@4	Margin	6.87	4.65	32.3%
pass@4	Five-signal block	6.84	4.76	30.3%

Table 16: Audit adds information beyond confidence signals in 74 majority-reward runs. Entries are in-sample R^2 for nested regressions on held-out gains. The five-signal block contains margin, plurality share, distinct-answer count, entropy and the identical-answer-group fraction.

	pass@1 gain		pass@4 gain	
Signals	Before audit	With audit	Before audit	With audit
Margin	0.430	0.846	0.337	0.813
Five-signal block	0.523	0.854	0.414	0.816

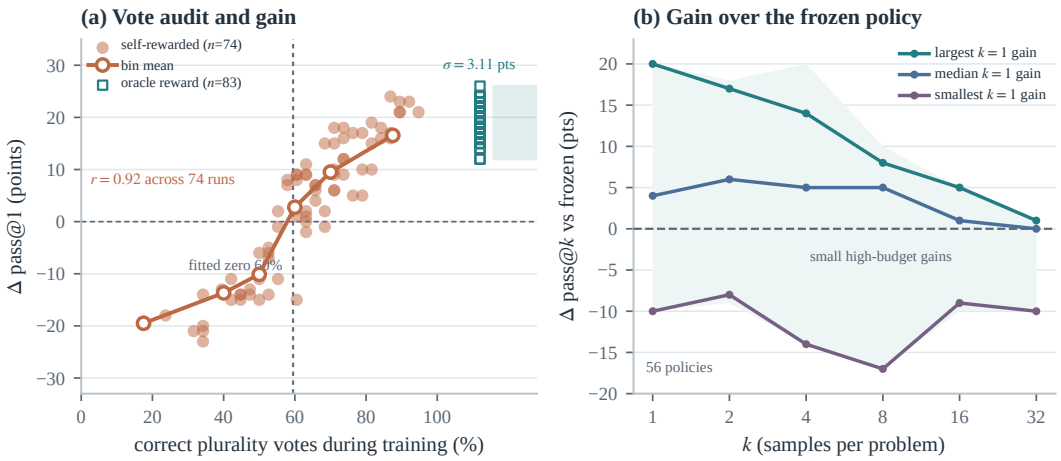

Figure 5: (a) Training-time vote correctness and held-out gain for 74 majority-rewarded runs, with binned means. Oracle runs occupy a separate display band; their horizontal position does not encode a measured vote audit. The dashed vertical line is the fitted majority-arm zero crossing, $A = 0.595$. (b) Coverage deltas for 56 trained policies against the frozen policy, showing the full range and three representative curves. The 32 policies with pass@1 gain above two points have pass@32 gains between zero and one point.

G Supplementary Reward-Channel Instruments

G.1 Finite-budget sampling

Gains concentrate at small sampling budgets. The coverage evaluation generates 32 samples per held-out problem. A representative majority-rewarded policy increases pass@1 from 0.34 to 0.54, while pass@32 stays at 0.97. An oracle policy reaches 0.52 and 0.98, respectively (table 10). Across the 32 policies with pass@1 gains above two points, the median gain changes from eight points at $k = 1$ to zero at $k = 32$ (table 11). This shows improved access to correct answers at a small sampling budget within the screened population. The full set of 56 policies is shown in fig. 5b.

Each retained problem had a correct base sample during screening, and the frozen policy already solves 97 of 100 evaluation problems at $k = 32$. Within this population, the curves measure how efficiently adaptation brings correct responses into a smaller sampling budget.

G.2 Training-Stream Perturbations

A label-blind intervention on input form. An adversary can modify a training prompt before the victim samples its vote group. We study bounded surface edits selected without querying labels, verifiers or gradients. The implementation enumerates numeric restatements, unit conversions, coreference substitutions and lexical edits, taking the first candidate within a similarity budget. With the default token-set Jaccard similarity, a candidate must satisfy $\mathrm{Sim}(x, x') \geq 0.8$. Evaluation questions remain unchanged. On the fixed GSM8K split this procedure rewrites 109 of 200 training questions: 84 lexical edits and 25 unit conversions. The gold annotation is carried over. These are input-form stress tests; lexical similarity by itself does not certify semantic equivalence (section D).

The audit connects perturbations to their training outcomes. Across 23 paired clean/perturbed cells, all settings except the training prompts are shared. The audit difference between the two training trajectories correlates with their held-out gain difference at $r = 0.933$, with slope 0.687 (fig. 6b). This is close to the majority-arm slope of 0.729. In the 17 cells where vote correctness is lower, its mean change is -0.116 and the mean gain difference is -0.070. In the six cells with higher vote correctness, the corresponding changes are $+0.162$ and $+0.113$. Surface edits can raise or lower vote correctness. In both groups, the mean held-out gain changes in the same direction as the audit.

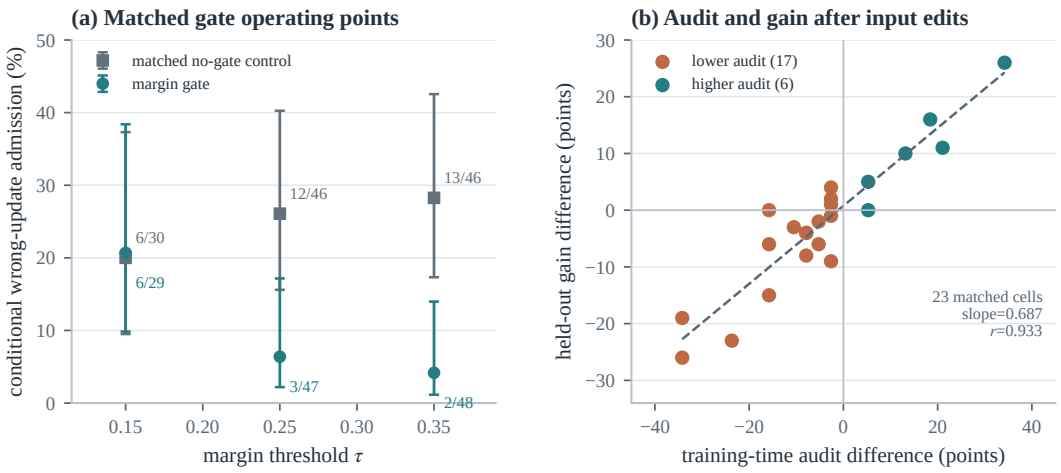

Figure 6: (a) Margin-gate operating points on synthetic arithmetic streams, with each gate compared to its own no-gate cells. Thresholds 0.25 and 0.35 implement the same rule at $N = 4$ and use different cells. Counts are accepted wrong votes among clean-correct vote attempts; error bars are descriptive Wilson 95% intervals. (b) Differences in training-time audit and held-out gain for 23 paired GSM8K perturbation cells. The fitted slope is 0.687, compared with 0.729 in the majority-reward study.

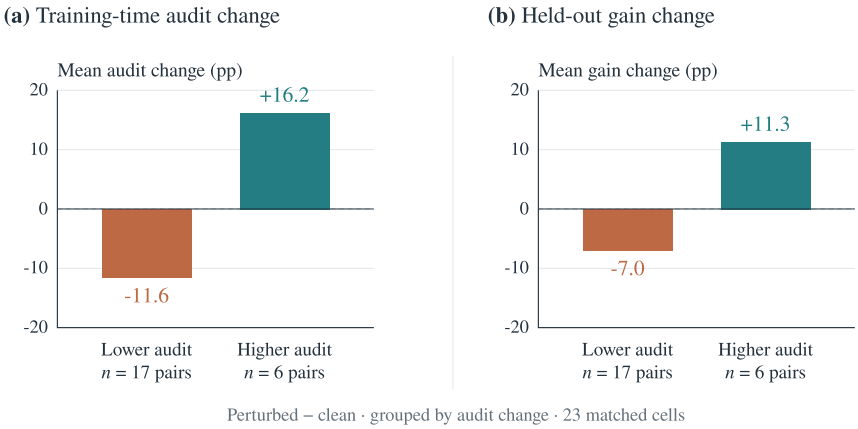

Figure 7: Input perturbations connect changes in supervision to changes in adaptation. Groups are defined by the sign of the observed audit difference; bars show means over all 23 matched cells. This grouping is based on observed outcomes rather than assigned intervention strengths. The individual paired observations appear in fig. 6b.

Two interventions expose the same empirical relationship. The reward-mixture experiment changes the target incentives directly. The input experiment leaves the majority-reward rule fixed and changes the prompts from which its targets are generated. Figure 7 shows that the two signs of the audit shift correspond to the two signs of the mean outcome change. On a common percentage-point scale, ten points of audit variation correspond to 7.29 points of fitted gain in the majority study and 6.87 points in the paired perturbation study. The similar slopes link reward correctness to adaptation outcomes in both settings. The audit thus provides a common way to describe how adaptation responds to changes in reward construction and training inputs.

G.3 Filtering Wrong-Vote Updates

Matched empirical operating points. We evaluate a margin gate on Llama-3-8B-Instruct in a separate streaming experiment with synthetic multiple-choice addition problems. Each branch processes the same 32-item stream, perturbing the first 16 items, and samples four responses per item.

Table 17: Matched 32-item streaming cells. Each row compares a margin threshold with its own no-gate branches. Acceptance is admitted updates divided by stream length; clean Δ reports held-out sample-accuracy changes for both branches. CWAR is defined by (9).

τ	Cells	No-gate CWAR	Gate CWAR	Accept	Δ no gate	Δ gate
0.15	2	6/30 (20.0%)	6/29 (20.7%)	78.1%	−0.047	−0.023
0.25	3	12/46 (26.1%)	3/47 (6.4%)	54.2%	−0.010	+0.010
0.35	3	13/46 (28.3%)	2/48 (4.2%)	47.9%	+0.005	0.000

The stream uses a centered group-reward policy-gradient update on LoRA parameters, as specified in section E. We record whether each proposed update uses a wrong pseudo-label and whether the gate admits it.

Let C_i denote that the clean-prompt plurality is correct immediately before the attempted perturbed update, and W_i that the perturbed rollout has a wrong plurality. The reported conditional wrong-update admission rate is

$$\mathrm{CWAR} = \frac{\sum_{i \in \mathcal{P}} \mathbf{1}[C_i \wedge W_i \wedge \mathrm{accepted}_i]}{\sum_{i \in \mathcal{P}} \mathbf{1}[C_i]}, \tag{9}$$

where $\mathcal{P}$ is the perturbed prefix. The label is the one used for the update, not a new sample after that update. The gate accepts when $d_i = (c_1 - c_2)/N > \tau$. Clean held-out sample accuracy is measured separately before and after processing the stream.

At $\tau = 0.35$, conditional wrong-update admission falls from $13/46$ to $2/48$, while the gate processes 46 of 96 stream updates and preserves mean held-out sample accuracy. At $\tau = 0.25$, it processes 52 of 96 updates and reduces admission from $12/46$ to $3/47$ (table 17). Each of the three cells at either threshold has a lower rate than its matched control. At $\tau = 0.35$, the decrease in CWAR is 24.1 percentage points, or 85% relative to the matched no-gate branches. Reporting admission alongside update volume and separately measured clean accuracy makes the trade-off explicit: the gate filters wrong supervision while continuing to admit updates. At $N = 4$, margins lie on a quarter-step grid, so thresholds 0.25 and 0.35 select the same possible margin values under the strict comparison; their rows summarize different matched cells. Appendix E.1 analyzes the finite-audit calibration rule.

H REWARD-SELECTION VALIDATION

Table 18: All observation windows and accuracy controls. Leave-one-seed-out RMSE in percentage points; each arrow adds selection correctness A. Mixture models control reward weight.

Cohort / target	Controls	10 groups	19 groups	38 groups
Majority / pass@1	P	$11.17 \to 9.20$	$7.98 \to 6.45$	$5.57 \to 5.14$
Majority / pass@1	P, P^2	$10.62 \to 9.17$	$7.69 \to 6.50$	$5.44 \to 5.14$
Majority / pass@1	P + five signals	$10.14 \to 9.56$	$7.37 \to 6.78$	$5.23 \to 5.20$
Majority / pass@4	P	$11.64 \to 9.78$	$8.77 \to 7.34$	$6.67 \to 6.06$
Majority / pass@4	P, P^2	$11.13 \to 9.76$	$8.16 \to 7.20$	$5.71 \to 5.44$
Majority / pass@4	P + five signals	$10.64 \to 10.17$	$8.17 \to 7.63$	$6.22 \to 6.17$
Mixture / pass@1	P	$7.97 \to 7.23$	$7.14 \to 5.99$	$5.10 \to 4.53$
Mixture / pass@1	P, P^2	$8.01 \to 7.24$	$7.11 \to 6.03$	$4.53 \to 4.22$
Mixture / pass@1	P + five signals	$7.65 \to 7.44$	$6.30 \to 6.01$	$4.68 \to 4.58$
Mixture / pass@4	P	$7.44 \to 6.86$	$6.90 \to 5.92$	$5.30 \to 4.74$
Mixture / pass@4	P, P^2	$7.47 \to 6.88$	$6.76 \to 5.94$	$4.07 \to 3.84$
Mixture / pass@4	P + five signals	$7.17 \to 7.04$	$5.97 \to 5.83$	$4.52 \to 4.52$

Analysis protocol. The reward-selection analysis is retrospective and uses the original 74 majority runs and 131 mixture records from 119 seeds. The three observation windows are the first 10 groups, the first 19 groups, and all 38 groups in callback append order. Each feature uses the same window. Response accuracy is the number of gold-matching responses divided by 8 times the number of groups; unparseable responses remain in the denominator. All OLS regressions include an intercept. Mixture fits include categorical indicators for weights 0.50 and 0.75, with 0.25 as reference. Every record sharing the held-out seed is excluded from fitting. The analysis reports all three windows and both pass@1 and pass@4 outcomes, with no window selection inside a fold. The outcome is the final gain over the frozen policy, not the gain after a measured intermediate checkpoint.

Uncertainty and fold sensitivity. For each nested comparison, we keep the out-of-fold predictions fixed and resample training-seed clusters 10,000 times. The intervals describe the change in RMSE over these errors; they do not refit the regression or supply independent confirmation of the exploratory analysis. With 19 groups, the pass@1 RMSE improvement over response accuracy has intervals $[0.54, 2.52]$ points for majority and $[0.51, 1.72]$ for mixtures. For the expanded accuracy-plus-concentration baseline these are $[-0.06, 1.24]$ and $[-0.02, 0.58]$ points. Twenty five-fold partitions use random seeds 7300–7319 and keep common training seeds together. Mean accuracy-only RMSE changes from 8.00 to 6.47 points for majority and 7.14 to 6.01 for mixtures; the expanded-control means change from 7.54 to 6.90 and from 6.32 to 6.05 points.