# WHEN REVISION BACKFIRES: CONSEQUENCE-AWARE VERIFICATION FOR MULTI-ROLE LLM REASONING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Verifier-guided reasoning is usually drawn as a monotone pipeline: a solver proposes, a verifier checks, a reviser repairs. That picture hides an asymmetry. Sending an answer back for revision is an action, not a prediction: replacing a correct answer with a wrong revision turns a point into zero, whereas keeping a wrong one merely preserves the status quo. Across 18 long multi-role math runs, unnecessary revision, not verdict error in general, is the failure most strongly tied to lost end-to-end value, negative within every run. We recast verification as consequence-aware control and introduce KEEPER, a preserve-first policy family that routes only what is worth rewriting and restores the solver answer when evidence for replacing it is weak. On a held-out controller panel, KEEPER improves three outcomes by 12.3–14.9 points and turns 17 of 19 repair opportunities into useful actions while degrading none. With the first-pass solver and reviser fixed, it beats a shared checker in all six comparisons while spending 13.5% fewer response tokens than always revising. A same-scale 7B preserve-first student, frozen and preregistered, improves all 12 checkpoints trained after it was frozen and 29 of 35 workflows of a problem-disjoint split with six unseen task families, blocking 99% of harmful revisions; a same-data variant improves every one of those 35 workflows over its gate. Score verifiers by the consequences of the routes they trigger, not by verdicts in isolation.

## 1 INTRODUCTION

Verifier-guided reasoning is usually described as a monotone improvement pipeline: a solver proposes an answer, a verifier checks it, and a reviser repairs rejected answers. This pattern appears in reward-model reranking, self-refinement, and multi-agent LLM systems (Cobbe et al., 2021; Lightman et al., 2024; Madaan et al., 2023; Wu et al., 2024). The description hides a control decision. Once rejection overwrites the solver answer, a verifier does not merely predict correctness; it decides whether another generation is worth the risk.

That risk is asymmetric. Rejecting a correct answer can replace it with a wrong revision, whereas accepting a wrong answer simply preserves the solver baseline. In long multi-role training runs, we observe that false rejection, not conventional verifier accuracy alone, tracks lost end-to-end value. This motivates a preserve-first controller that asks a counterfactual question: *will revision improve the answer already in hand?*

The distinction is easiest to see on two superficially identical verifier errors. If the solver is correct and the reviser is wrong, an unnecessary rejection changes one system point into zero. If both answers are wrong, an unnecessary acceptance leaves system accuracy unchanged. A classifier may penalize both decisions, but deployment does not. Conversely, identifying a wrong solver answer helps only when the available reviser can actually repair it. A verifier should therefore be evaluated, and its decisions made, by the *consequence of the route*, not by a verdict considered in isolation.

We introduce KEEPER, a false-rejection-aware controller for solver–verifier–reviser systems. We instantiate it at two decision points: pre-revision routing, which can avoid a revision, and post-generation selection, where a semantic-and-support policy chooses among already generated candidates and a preserve rule may restore the solver answer. LoRA is not our starting assumption: we



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.

first choose a multi-role design, then use role-specific low-rank adapters over a frozen backbone so that checker and revision updates need not store several full models (Hu et al., 2022).

Our evidence separates four questions. An audit of 18 long math runs tests whether false rejection is associated with system gain. A fixed 114-case role-action intervention tests whether targeted preservation and repair change outcomes. A pipeline comparison fixes the first-pass solver and reviser adapter and changes only the checker, and candidate replay then holds the generated alternatives fixed and matches revision counts, isolating the value of choosing which examples to route. Finally, a same-scale 7B preserve-first student is frozen and evaluated, under preregistered rules, on checkpoints trained after it was frozen and on a problem-disjoint split with task families it never saw. Across these, the result is consistent: when revision can both repair and damage answers, preserve-first control improves the final system.

**Key results.** A held-out controller turns 17 of 19 repair opportunities into useful actions and degrades none. With solver and reviser fixed, specialized routing beats a shared checker in all six comparisons at 13.5% fewer response tokens. A frozen 7B preserve-first student blocks 99% of harmful revisions on 184 problems it never saw, and a same-data variant improves all 35 workflows over their gates.

**Contributions.**

- We formulate verification as a cost-sensitive accept-or-revise decision and decompose system gain exactly into beneficial and harmful revisions.
- We introduce KEEPER, a preserve-first controller that acts before revision and after generation, and a same-scale 7B instantiation of it.
- We show gains in a held-out intervention, in six fixed-solver pipeline comparisons against a shared checker, and in preregistered transfer to new checkpoints and to problems and task families never seen in training.

## 2 VERIFICATION AS CONTROL

### 2.1 GATED SYSTEM VALUE

For input $x$, let $S, R \in \{0, 1\}$ denote correctness of the solver answer $a_s$ and revision $a_r$. Let $V = 1$ mean "accept the solver" and $V = 0$ mean "return the revision", so final correctness is

$$Y = VS + (1 - V)R. \tag{1}$$

The gain over returning the solver is

$$\mathrm{SG} = \mathbb{E}[Y - S] = \mathbb{E}[(1 - V)(R - S)]. \tag{2}$$

Revision is beneficial only on $S = 0, R = 1$ and harmful on $S = 1, R = 0$. A correctness classifier is therefore insufficient: two wrong answers can need different actions if only one is repairable.

Let $z$ denote deployment-visible information and let $q(z)$ be the predicted probability that the current answer should be preserved. A consequence-weighted objective weights harmful revision and missed repair separately:

$$\begin{aligned} \mathcal{L}_{\mathrm{gate}} = & -\lambda_{\mathrm{harm}} \mathbf{1}[S{=}1, R{=}0] \log q \\ & -\lambda_{\mathrm{miss}} \mathbf{1}[S{=}0, R{=}1] \log(1 - q). \end{aligned} \tag{3}$$

Neutral cases ($S = R$) affect neither this loss nor Eq. 2. The decisive-only 7B preserve students (Section 4.3) minimize it directly, training only on class-balanced rows with $S \neq R$. The checkers in the six pipeline comparisons receive its policy-gradient form: their reward is final correctness $Y$, so rejecting rather than accepting changes it by $R - S$; every training episode generates the revision whatever the verdict, so $R$ is observed for accepted answers too. Routing sees only $z$.

### 2.2 A CONSEQUENCE TAXONOMY

The four $(S, R)$ combinations expose why ordinary verifier accuracy is an incomplete objective (Figure 1). When $S = R$, routing cannot change correctness. When $S = 0, R = 1$, rejection

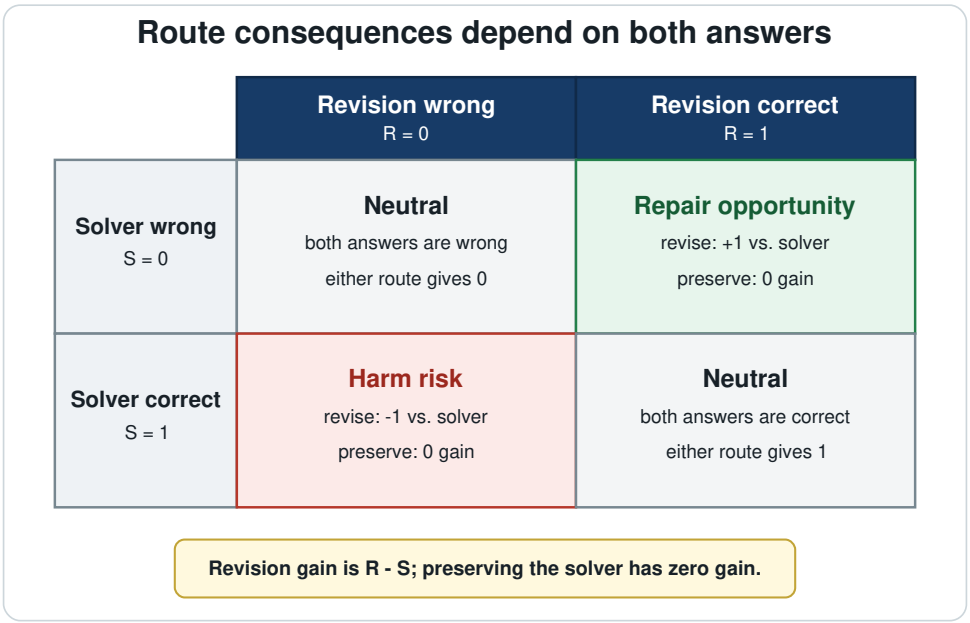


Figure 1: Route consequence depends jointly on solver correctness $S$ and available-revision correctness $R$: routing is correctness-neutral when $S = R$, beneficial at $(0, 1)$, and harmful at $(1, 0)$.

realizes a repair opportunity; when $S = 1, R = 0$, the same action destroys a correct answer. Define the label-aware route value as $Y^\star = \max(S, R)$. Let $\pi_{\mathrm{FA}} = \Pr(V{=}1, S{=}0)$ and $\pi_{\mathrm{FR}} = \Pr(V{=}0, S{=}1)$ denote false-accept and false-reject event probabilities, and let $\rho_{\mathrm{rep}} = \Pr(R{=}1 \mid V{=}1, S{=}0)$ and $\rho_{\mathrm{harm}} = \Pr(R{=}0 \mid V{=}0, S{=}1)$. Then

$$\mathbb{E}[Y - Y^\star] = -\pi_{\mathrm{FA}}\rho_{\mathrm{rep}} - \pi_{\mathrm{FR}}\rho_{\mathrm{harm}}, \tag{4}$$

an exact decomposition for binary correctness. An unparsable verdict is first mapped to its documented fallback action $V$ and then counted in the corresponding event. The identity does not assume that one consequence is universally larger; it requires measuring both event prevalence and route outcome rather than optimizing a symmetric verdict score.

This view also clarifies why sparse end-to-end reward is difficult to interpret. The same terminal zero may come from an unsolved problem, a missed repair, an unparsable verdict, or a verifier-induced regression. A role-specific controller can expose these events directly instead of asking one delayed bit to teach discrimination, calibration, formatting, and routing at once.

### 2.3 KEEPER DECISION RULE

Two interfaces instantiate the same preserve-first principle (Figure 2). In *pre-revision routing*, a checker decides whether to invoke one revision; this is the interface used in the six pipeline comparisons and token accounting. In *post-generation selection*, candidates already exist; a semantic-and-support selector chooses among them and a preserve rule may restore $a_s$. The latter is used only in the 458-example diagnostic and does not itself save candidate-generation compute. We report the two interfaces separately and do not pool their effects.

In the post-generation interface, a semantic judge scores candidate correctness, a candidate module records answer support, and the preserve rule falls back to $a_s$ when the selected candidate has both a low score and low support. The cross-fitted rule is

$$\begin{aligned} \text{fallback to } a_s \quad \text{if} \quad & q_{\mathrm{select}} < 0.55 \\ & \wedge \ \mathrm{support} < 2. \end{aligned} \tag{5}$$

The rule family and thresholds are selected independently inside each training fold and applied to its held-out fold. All five folds select the same rule. This design makes rejection conditional on positive evidence rather than the absence of confidence.

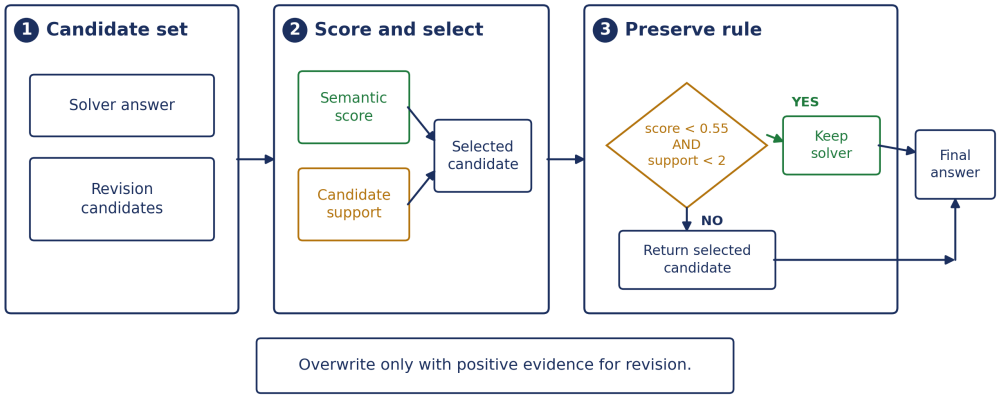


Figure 2: KEEPER combines a semantic score with candidate support and preserves the solver answer when both replacement signals are weak. The shown threshold is the out-of-fold rule selected in the reused 458-example diagnostic; labels select rules on training folds but are not decision inputs.

**Why agreement is not a second verifier.** Semantic judgment and agreement answer different questions. The former asks whether a candidate is compatible with the problem and its reasoning; the latter asks whether independently generated paths converge on the same answer (Wang et al., 2023). Agreement alone can amplify a shared misconception, while a single semantic score can be brittle to wording. Either signal alone can justify replacing the solver answer; the rule restores that answer only when both are weak, so a repair detected by one signal is not discarded.

**Training and deployment separation.** During development, executable or exact-match checks reveal $S$ and $R$ and therefore identify beneficial, harmful, and neutral routes. These labels train or select the controller but are not inputs to it. At deployment, KEEPER receives the problem, generated text, semantic scores, and candidate agreement only. In the 458-example diagnostic, both the base selector (one of 25 frozen policies) and the preserve rule can be chosen out of fold; a nested cross-fit that selects both on the other four folds picks the same pair in every fold, so the selector is out of fold.

### 2.4 ROLE ADAPTATION WITH LORA

The solver, checker, and reviser use role adapters over a frozen Qwen2.5-7B-Instruct backbone (Qwen Team et al., 2025). Each role has a low-rank update $\Delta W_\tau = B_\tau A_\tau$. This choice preserves a common capability base while exposing compact role-owned modules: the checker can be calibrated without rewriting the solver, and a reviser update can be evaluated under a fixed gate. The controller itself is independent of LoRA and applies to any implementation that exposes the same answers and decision scores. The 458-example candidate diagnostic is an explicit exception to the shared-backbone setup: its external `gpt-oss-120b` judge makes it neither same-model nor matched-compute evidence.

## 3 EVALUATION PROTOCOL

**Evidence tiers.** The long-run audit contains 4,186 logged snapshots from 18 math-reasoning runs trained on a local subset of DeepMath-103K (He et al., 2025): long non-pretrained runs from one training campaign with at least 194 logged rows. In the broader frozen 75-run, 6,570-row universe the association is also negative ($r = -0.151$). The held-out controller panel contains 114 paired role-action cases and tests a fixed counterfactual behavior. The candidate diagnostic contains 458 examples partitioned into five deterministic source-stratified folds. Candidates and scores are fixed, but the semantic-and-support base policy was selected from 25 frozen policies using labels from the full panel. The selected policy ranks exact-output groups by semantic score plus $0.1 \log(\text{support})$,

with an invalid-output fallback to the designated first-pass solver answer. The policy was first chosen on the full panel; re-selecting it and the preserve rule jointly inside a nested five-fold cross-fit chooses the same policy and rule in all five folds (Appendix C), so the full $+16.81$-point system gain is out of fold. The panel combines 30 AIME 2024 and 30 AIME 2025 problems obtained from the Maxwell-Jia and FVU mirrors of the MAA competition (Mathematical Association of America, n.d.; Maxwell-Jia, 2024; FVU, 2025), a 200-question subset of the `cais/mmlu` validation split (Hendrycks et al., 2021), and all 198 GPQA Diamond questions (Rein et al., 2024).

**Held-out role–action controller.** The 114-case experiment evaluates a rank-64 Qwen2.5-7B LoRA controller trained on a disjoint 32,000-row selector set. For each fixed workflow state, it jointly scores view–action pairs using only five counterfactual role-view identifiers, their gate and selector decisions, and their score margins; hidden consequence fields are unavailable at inference. Its selected action agrees with the action attached to its selected view in all 114 cases. The original-view and selected-view outcomes therefore reuse the same state and candidate pool. A provenance audit found zero train/evaluation overlap by row ID, source-state ID, source-row ID, or exact prompt hash.

**Controlled comparisons.** We freeze the first-pass solver and reviser adapter and change the checker. Because checker feedback conditions revision, this comparison measures the complete checker-specialized pipeline, not selection alone. We evaluate two independent checkpoints on three panels where revision can both repair and damage answers: sparse-clue ($n = 768$), strict-no-position ($n = 768$), and constraint-stress ($n = 1{,}536$). Related templates are resampled as groups when computing 95% intervals. A stricter replay compares against outcome-blind random selection that matches the learned policy's revision count within each task family and reuses the same candidates, isolating selection; its null distribution is enumerated exactly rather than sampled.

**Baselines and metrics.** Baselines are solver-only, always revise, a shared generative checker, a solver-only rank-64 LoRA with the same ten-update budget, score thresholds with and without Platt calibration, PAG (Jiang et al., 2025), and outcome-blind selectors. We report final accuracy, gain over the solver, false rejection, revision rate, response tokens, paired intervals, and exact random-selection tail probabilities. In the 114-case panel, *useful* means a correct terminal action or a feasible REDO with a future correct candidate; *terminal* means a correct STOP or visible-candidate selection; and *exact action match* requires the outcome-derived target action and, for selection, target slot.

**Implementation and evidence boundaries.** The learned roles use rank-64 adapters on Qwen2.5-7B-Instruct. The two compact training replicates run ten updates each at learning rate $2.25 \times 10^{-6}$, covering 320 episodes and 428,917 prompt-plus-response tokens in total; each uses one H200 and together they take 0.584 measured accelerator-hours, a total that covers only these two runs (see Limitations). For the separate 458-example diagnostic, deterministic source hashing defines five folds; each fold searches the same family of 265 score-and-support rules on the other four folds. Its `gpt-oss-120b` semantic scores and fixed candidate frontier were computed before this analysis but are not compute matched to the 7B system. An anonymous standard-library replay checks the retained aggregate summaries in Appendix J, including paired effects, fold totals, randomization bounds, and compute arithmetic.

**Leakage controls and reproducibility.** The preserve-rule analysis uses deterministic source-stratified folds, and each fold decision comes from a rule selected on the other four folds; a nested variant also re-selects the base policy per fold and leaves the result unchanged. Pipeline comparisons fix the first-pass solver and reviser adapter; fixed-candidate replay additionally reuses generated outputs across selectors. Bootstrap units are source groups rather than individual near-duplicate templates. Random selectors are drawn independently within task family and constrained to the learned policy's revision count. These choices preserve the pairing that makes the contrast interpretable and prevent a task-mix shift from masquerading as better routing.

Table 1: Descriptive association over 18 long math runs (4,186 snapshots). Left: pooled correlation of each logged verifier statistic with system gain. Right: robustness of the false-rejection association.

| Statistic | $r$ | False rejection | Value |
|---|---|---|---|
| False rejection | $\mathbf{-0.259}$ | Run-cluster 95% CI | $[-0.366, -0.184]$ |
| False acceptance | $+0.030$ | Run-demeaned | $r = -0.303$ |
| Acceptance rate | $+0.108$ | First differences | $r = -0.423$ |
| Unparsable verdict | $+0.025$ | Leave-one-run-out | $[-0.294, -0.228]$ |
| Solver accuracy | $-0.011$ | Within-run sign | 18/18 negative |

Table 2: Paired outcomes on 114 held-out fixed cases (49 source episodes); intervals resample episodes.

| Metric | Original | Selected | $\Delta$ [95% interval] |
|---|---|---|---|
| Useful output | 50.88 | 65.79 | $+14.91$ $[+7.83, +22.88]$ |
| Terminal correct | 13.16 | 25.44 | $+12.28$ $[+5.93, +19.67]$ |
| Exact action match | 37.72 | 50.88 | $+13.16$ $[+5.41, +21.67]$ |

# 4 Results

## 4.1 False Rejection Tracks Lost Value

Table 1 reconstructs the long-run association. Of the logged verifier statistics, only false rejection is materially correlated with system gain ($r = -0.259$; false acceptance $+0.030$, unparsable verdicts $+0.025$); the relationship remains negative after run demeaning, first differencing, and leave-one-run-out analysis. These dependent snapshots do not establish causality, but they identify the failure mode we target.

Adjacent snapshots within a run are not independent. Run demeaning removes between-run offsets, first differences ask whether changes in false rejection move with changes in system gain, and leave-one-run-out analysis checks whether one trajectory determines the sign. All four agree, and the paired comparisons below test the same target interventionally.

## 4.2 A Held-Out Fixed-Panel Controller Improves Outcomes

On the 114 paired cases, the controller-selected view improves all three reported outcomes (Table 2). The cases come from 49 source episodes, so intervals resample whole episodes: terminal correctness, which credits no future repair, improves in 11 episodes and worsens in none. The largest useful gain occurs on the 19 repair opportunities: 17 improve and none degrade.

The aggregate improvement is concentrated where intervention is supposed to matter. Seventeen of 19 repair opportunities improve and none degrade. Of 58 already-useful originals, 57 remain useful and the single oracle-conflict case loses useful status. This decomposition distinguishes a controller that selects valuable actions from one that merely changes many actions.

## 4.3 Cross-Fitted Preserve-Rule Diagnostic

The fixed-candidate panel in Table 3 separates semantic candidate selection from the preserve rule. The semantic-and-support base policy, one of 25 frozen policies, raises accuracy from 42.58% to 58.52%; its external `gpt-oss-120b` score is combined with log support. When both the base policy and the preserve rule are selected on the other four folds, all five folds choose the same pair, so the whole selector is out of fold. Out-of-fold fallback changes 18 decisions: four are beneficial, none harmful, and 14 correctness-neutral. The preserve rule therefore adds 0.87 points (item bootstrap $[+0.22, +1.75]$; exact one-sided sign $p = 0.0625$ on the four decisive switches), reaching 59.39% ($[54.80, 63.76]$), and reduces false rejection from 9/458 to 5/458. The 120B judge is not essential. A Qwen2.5-7B LoRA student of its verdicts agrees with the teacher on 99.1% of identity-held-out items, where the untuned base emits almost no parseable verdict (Appendix D). As a preserve-first veto on a concurrent routing panel, the same-scale rule raises collapsing Shared workflows by 13–17 points (Appendix F), and a 7B student trained to preserve unless revision is strictly better improves

Table 3: Component accounting on a reused 458-example diagnostic panel. A nested cross-fit re-selects the base policy and the preserve rule per fold and chooses the same pair in all five folds, so every row is out of fold. Accuracy and FR are percentages; deltas are points. Semantic scores use `gpt-oss-120b` and are not compute matched.

| Policy | Correct | Acc. | Δ solver | Δ prior | FR |
|---|---|---|---|---|---|
| Solver | 195 | 42.58 | — | — | — |
| Semantic + support | 268 | 58.52 | +15.94 | +15.94 | 1.97 |
| **Base + rule** | **272** | **59.39** | **+16.81** | **+0.87** | **1.09** |

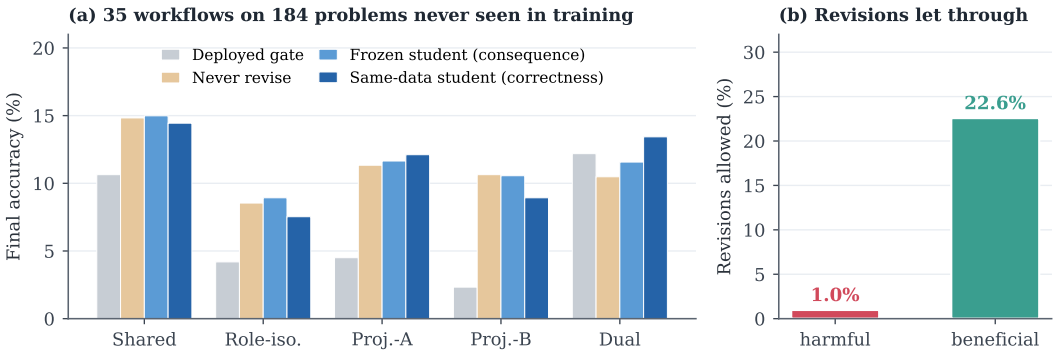


Figure 3: Preregistered transfer of frozen 7B preserve-first students to 184 problems never seen in training, across 35 routing workflows. (a) Mean final accuracy per training variant. (b) Of the gates' revisions, the frozen student lets through 1.0% of the harmful ones and 22.6% of the beneficial ones.

all eight held-out-problem panels of seeds 45–48 without any threshold (Shared +15.71, Dual +1.99 points; Appendix G). Frozen and preregistered, it improves all 12 checkpoints trained after it was frozen (Appendix G.2) and 29 of 35 workflows of a problem-disjoint split with six task families it never saw, where it blocks 403 of the gates' 407 harmful revisions and beats a random veto of the same size on 34 of 35 workflows (Figure 3; Appendix G.4). A preregistered same-data student trained on correctness labels, under the same preserve-first rule, improves all 35 workflows of that split over their gates and beats both the gate and never revising on every Dual workflow, whose revisions already help (Appendix G.3).

### 4.4 Checker Specialization and Selection

With first-pass solver and reviser adapter fixed, specialized checking improves final accuracy over a shared generative checker in all six comparisons (Figure 4). Every grouped 95% interval is above zero and paired $p \leq 0.015$. Because feedback changes the reviser's conditioned output, this is a pipeline-level specialization effect. The fixed-candidate replay provides the narrower selection result: matching revisions separately within each task family leaves a +2.558 to +4.168 pp advantage over random selection, with an exact tail probability below $2 \times 10^{-7}$ in each comparison.

### 4.5 Baselines and Cost

Specialized routing exceeds always revise in all six cells while using 13.49% fewer response tokens and 11.8% fewer model calls. Family-matched random selectors trail by +2.558–+4.168 pp (exact tail $< 2 \times 10^{-7}$ in every comparison), and two short training seeds remain positive (+3.731 and +4.295 pp). On a separate 1,578-example threshold diagnostic, a scalar score needs a 73.38% revision rate to match Keeper's 29.21% accuracy, versus 60.71% for the consequence-aware rule; at the matched budget, scalar rules reach 27.69–28.01%. Against the released PAG profile on 768 matched questions, Keeper is 4.56 points more accurate while using 58% of its response tokens; a PAG profile with a stronger domain-neutral prompt is 4.56 points more accurate than Keeper, and neither comparison matches total supervision (Appendix I).

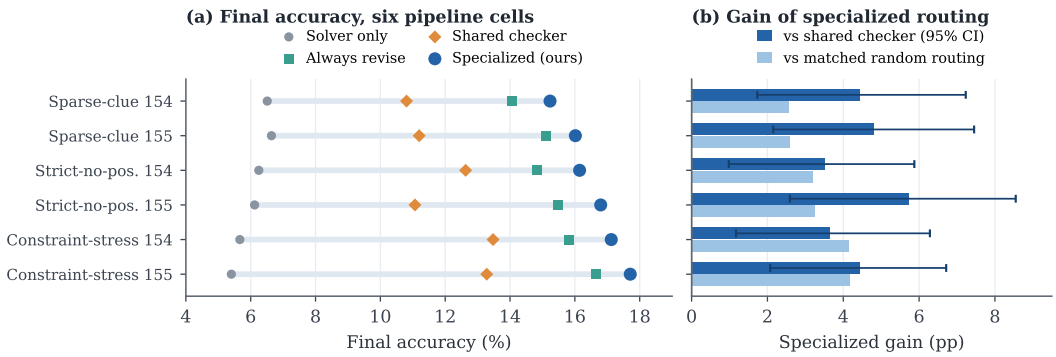


Figure 4: Six fixed-solver pipeline comparisons (two training replicates on three panels). (a) Final accuracy of the first answer, always revising, a shared checker, and the specialized checker. (b) Gain of specialized routing over the shared checker (95% template-group interval) and over random routing matched to its revision count within each task family (Table 23).

Table 4: Same-scale 7B logistic preserve rule on 1,472 items × 3 replicates. The rule is fit on solver/checker text only and never adds a revision. Values are percentages except Δ (points).

| Training variant | Current | Preserve | Δ | FR 0→1 |
|---|---|---|---|---|
| Dual read/write | 28.42 | 29.35 | +0.93 | 8.38→2.42 |
| Shared adapter | 26.92 | 46.78 | +19.86 | 39.49→1.22 |
| Role-isolated | 7.84 | 28.37 | +20.54 | 21.90→0.14 |
| Projected-gradient A | 4.94 | 30.82 | +25.88 | 29.30→0.02 |
| Projected-gradient B | 0.34 | 34.90 | +34.56 | 34.71→0.00 |

The 458-example preserve rule above uses an external larger judge. A same-scale controller can be fit on the 7B solver and checker text of a concurrent routing panel (Anonymous, 2027): 1,472 items under three training replicates, five-fold identity-grouped cross-fitting, and a preserve-first veto that never adds a revision. Table 4 shows that this 7B-text rule improves the pooled result of all five training variants (14 of 15 variant–replicate cells; Dual-44 moves −0.20) and recovers the most value from workflows whose revision stage had been destroying first answers. The same protocol on Dual/Shared seeds 45–48 repeats the pattern (Appendix E).

The two independent checkpoints improve within their own generated traces; all 12 intermediate measurements of those short runs are positive, at 0.584 H200 GPU-hours, 20 updates, 320 episodes, and 428,917 training tokens. This repeatability test shows that the routing effect survives fresh optimization noise at modest cost.

## 5 Analysis and Deployment Implications

### 5.1 When Preserve-First Control Helps

Solver-only avoids routing harm but discards every repair; always revise captures repairs but rewrites correct answers. Specialized pipelines beat their own first-pass answers by 8.72–12.31 points, and selective routing is 0.91–1.30 points better than always revise at lower token cost. A scalar correctness threshold ignores whether the available revision is better, and a global revision budget is the wrong control because families differ in repairability. Family-matched replay, fixing candidates, feedback, revision counts, and calls, still leaves 2.56–4.17 points from *which* examples it routes.

KEEPER is for the regime where both $\Pr(S{=}0, R{=}1)$ and $\Pr(S{=}1, R{=}0)$ are nonzero. The expected benefit of revising on a region $\mathcal{A}$ is

$$\Pr(S{=}0, R{=}1, x \in \mathcal{A}) - \Pr(S{=}1, R{=}0, x \in \mathcal{A}). \tag{6}$$

A useful gate makes the first term exceed the second. Agreement and semantic evidence are proxies for that inequality; false rejection is allowed only when expected repair value justifies the risk.

### 5.2 Monitoring, Errors, and Scope

Audit consequence, not verdict accuracy: solver and final accuracy, repair opportunity, harmful rewrite, false rejection, revision rate, parse failures, and tokens. A rising revision rate without final gain is a direct warning. The preserve rule can lower false rejection while raising false acceptance when both answers are wrong; reporting both errors prevents a favorable endpoint from hiding the trade-off. Before deployment, compare solver-only, always-revise, and family-matched random routing on fixed candidates, and require a gain in both value and the accuracy–cost trade-off.

The claim is a control claim for workflows in which revision can both repair and damage answers. It holds across evidence of different kinds: an audit of long training runs, a held-out intervention, fixed-solver pipeline comparisons, same-candidate replay, and preregistered transfer of a frozen same-scale rule to new checkpoints and new task families.

## 6 Related Work

**Verifiers and self-correction.** Outcome and process reward models score candidate solutions (Cobbe et al., 2021; Lightman et al., 2024; Wang et al., 2024; Zhang et al., 2024). Self-Refine, Reflexion, and self-critique generate feedback before another attempt (Madaan et al., 2023; Shinn et al., 2023; Saunders et al., 2022). PAG alternates a unified policy between generation and verification and selectively revises rejected attempts (Jiang et al., 2025). KEEPER asks a complementary question: should the answer in hand be replaced?

**Self-correction as feedback control.** Liu & Meng (2026) treat repeated intrinsic self-correction as two-state Markov dynamics. Their action is whether to continue same-model refinement. We instead route between an existing solver answer and revision candidates, decompose each route by $(S, R)$, and evaluate both pre-revision routing and a post-generation preserve rule.

**Cost-sensitive and multi-agent learning.** Asymmetric losses (Elkan, 2001; Lin et al., 2017) and multi-agent credit assignment (Foerster et al., 2018; Rashid et al., 2018) reweight errors or delayed outcomes. Equation 2 instead uses the workflow to name the samples where revision changes value.

**Training-time role routing.** Concurrent anonymous work studies read/write routing among role-specific LoRA adapters (Anonymous, 2027). That work changes how role parameters are trained; KEEPER changes the inference-time accept-or-revise decision, so the two contributions are distinct.

## 7 Conclusion

A verifier should not be rewarded merely for finding errors; it should be judged by the consequence of its decision, and deployed behind a rule that preserves the answer in hand unless revision has positive evidence. Same-candidate replay shows that which examples are revised matters beyond revision volume, and a frozen same-scale 7B preserve-first rule transfers to new checkpoints and to task families it never saw, blocking nearly every harmful revision. KEEPER turns "verify, then revise" into a consequence-aware rule: preserve unless revision has positive evidence.

## 8 Limitations

The long-run audit is associative and math-only. The 114-case intervention and the 458-example selection study use fixed candidate panels, and the latter's semantic scores come from a larger external judge, which the 7B student approximates. The pipeline comparisons use three repair-heavy panels and two training replicates, and checker feedback conditions the revision text, so they measure the complete checker-specialized pipeline. The same-scale preserve rule is evaluated on a concurrent routing panel and its development validation mix; broader domains and larger models remain to be tested. Reported accelerator time covers the compact training replicates only.

## Ethics Statement

Selective revision can reduce unnecessary generation, but it is not a safety guarantee: a mistaken verifier can preserve an incorrect or unsafe answer or replace a correct one. High-stakes use therefore requires domain-specific validation and human oversight. Our experiments use existing research benchmarks and model-generated candidates rather than newly recruited human participants. The source cards record MIT for DeepMath-103K, the AIME-2024 mirror, and MMLU, and CC BY 4.0 plus a no-online-example access condition for GPQA; the AIME-2025 mirror exposes no explicit license metadata. We therefore report only aggregate benchmark results and redistribute no task instances. The gpt-oss judge is Apache-2.0 and remains subject to its usage policy. We expose failure modes and compute-relevant components to help readers assess deployment risk. The aggregate evidence and original consistency code are distributed under Apache-2.0; we redistribute neither model weights nor raw benchmark instances. All authors have read and adhere to the ICLR Code of Ethics.

## Reproducibility Statement

The decision rule and its two instantiations are specified in Section 2, and every evaluation protocol, evidence tier, and leakage control is stated in Section 3. Appendix A records the estimand and inferential unit for each tier so that no result is read outside its scope. Appendix J gives the anonymous reproduction contract, listing the fixed panels, frozen candidate payloads, fold assignments, and aggregate metric definitions needed to recompute every reported number. Anonymized aggregate evidence, per-item preserve decisions, and the scripts that recompute them are provided as supplementary material.

## AI Use Statement

We used generative AI tools to assist with writing, coding, and data analysis. The authors reviewed all AI-assisted content and take full responsibility for this paper.

## References

Anonymous. Read together, write apart: Decoupling rollout access from update ownership in multi-role LLM training, 2027. Under review.

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

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 973–978, 2001.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 2974–2982, 2018.

FVU. AIME 2025 Dataset. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/FVU/AIME_2025.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. DeepMath-103K: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. arXiv preprint arXiv:2504.11456, 2025. URL https://arxiv.org/abs/2504.11456.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Yuhua Jiang, Yuwen Xiong, Yufeng Yuan, Chao Xin, Wenyuan Xu, Yu Yue, Qianchuan Zhao, and Lin Yan. PAG: Multi-turn reinforced LLM self-correction with policy as generative verifier, 2025. URL https://arxiv.org/abs/2506.10406.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Aofan Liu and Jingxiang Meng. Self-correction as feedback control: Error dynamics, stability thresholds, and prompt interventions in LLMs, 2026. URL https://arxiv.org/abs/2604.22273.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Mathematical Association of America. MAA Invitational Competitions: American Invitational Mathematics Examination. Competition website, n.d. URL https://maa.org/maa-invitational-competitions/. Accessed 2026-08-04.

Maxwell-Jia. AIME 2024 Dataset. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/Maxwell-Jia/AIME_2024.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5: Technical report. arXiv preprint arXiv:2412.15115, 2025. URL https://arxiv.org/abs/2412.15115.

Tabish Rashid, Mikayel Samvelyan, Christian Schröder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4295–4304. PMLR, 2018. URL https://proceedings.mlr.press/v80/rashid18a.html.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. arXiv preprint arXiv:2206.05802, 2022. URL https://arxiv.org/abs/2206.05802.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2303.11366.

Table 5: Scope and uncertainty of each evidence tier.

| Tier | Visible observations | Inferential unit | Supported conclusion |
|---|---|---|---|
| Long-run audit | 4,186 snapshots, 18 runs | Training run | Descriptive association |
| Role-action intervention | 114 cases, 49 episodes | Source episode | Fixed-panel action effect |
| Preserve-rule diagnostic | 458 examples, 5 folds | Held-out example | Out-of-fold preserve effect |
| Checker specialization | 6,144 evaluations | Template group | Pipeline checker effect |
| Matched-random replay | Panel–replicate cells | Revision family | Fixed-candidate selection |

Table 6: Intervention effects by counterfactual opportunity. Effects are percentage points; $+/0/-$ counts row-level useful-output changes.

| Opportunity slice | $n$ | Useful $\Delta$ | Terminal $\Delta$ | Exact $\Delta$ | Useful $+/0/-$ |
|---|---|---|---|---|---|
| Preserve useful original | 57 | $+0.00$ | $+1.75$ | $+5.26$ | $0/57/0$ |
| No useful oracle gain | 37 | $+2.70$ | $+0.00$ | $+8.11$ | $1/36/0$ |
| Repair opportunity | 19 | $+89.47$ | $+68.42$ | $+42.11$ | $17/2/0$ |
| Oracle conflict | 1 | $-100.00$ | $+0.00$ | $+100.00$ | $0/0/1$ |

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2203.11171.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Ahmed H. Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *First Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2308.08155.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. arXiv preprint arXiv:2408.15240, 2024. URL https://arxiv.org/abs/2408.15240.

## A Evidence Tiers and Uncertainty

Table 5 records the estimand and inferential unit for every evidence tier. We do not pool these datasets because repeated snapshots, paired comparisons, held-out preserve-rule predictions, and related structural templates have different dependence structures.

For the audit, we resample runs and check demeaning, first differences, and leave-one-run-out estimates. Intervention intervals resample the 49 source episodes, so repeated states of one episode move together, and the sign test counts episodes. Pipeline intervals resample structural template groups. The matched-random null is exact conditional on the number of revisions within each task family, rather than a Monte Carlo approximation.

## B Intervention Decomposition

The intervention gate agrees with 111/114 fixed labels (Wilson 95% interval $[92.55, 99.10]$), while the higher-level action family agrees on 110/114 ($[91.32, 98.63]$). These agreement scores describe the frozen diagnostic and are not external generalization estimates.

Table 7: Held-out preserve-rule results by source-stratified fold. Accuracy, gain, and false rejection (FR) are percentages or percentage points.

| Fold | $n$ | Switches | Beneficial | Harmful | Accuracy | Gain vs. solver | FR |
|---|---|---|---|---|---|---|---|
| 1 | 92 | 3 | 0 | 0 | 56.52 | 22.83 | 0.00 |
| 2 | 92 | 4 | 1 | 0 | 61.96 | 17.39 | 0.00 |
| 3 | 92 | 4 | 0 | 0 | 56.52 | 14.13 | 0.00 |
| 4 | 91 | 4 | 2 | 0 | 64.84 | 13.19 | 3.30 |
| 5 | 91 | 3 | 1 | 0 | 57.14 | 16.48 | 2.20 |
| Out of fold | 458 | 18 | 4 | 0 | 59.39 | 16.81 | 1.09 |

## C Cross-Fitted Preserve Rule by Fold

The 265-rule family is identical across folds. Every training split selects the score-below-0.55 and support-below-2 conjunction in Eq. 5. The base policy was originally chosen from the 25 frozen policies on the full panel. A nested cross-fit removes that step from the labels as well: in each fold both the base policy (maximizing training-fold final accuracy) and the preserve rule are selected on the other four folds only. All five folds choose the same policy and the same rule, so the nested out-of-fold result equals the one below exactly: 272/458 correct, system gain $+16.81$ pp (item bootstrap $[+13.32, +20.52]$). The same code with the base policy pinned reproduces the published cross-fit, which is the guard for the nested run. The deterministic fold-assignment SHA-256 begins `4a05a59c4ca2`; the accompanying decision-rule ledger records the full digest.

The semantic-and-support base has 268/458 correct, nine false rejects, and 137 false accepts. Adding the cross-fitted preserve rule yields 272/458 correct, five false rejects, and 151 false accepts. Four switches are beneficial, none harmful, and 14 neutral. Thus the preserve rule's incremental endpoint effect is $4/458 = 0.873$ points.

## D 7B Semantic-Verdict Student

Table 8 reports the compute-matched student of the `gpt-oss-120b` semantic labels used in the 458-example diagnostic. Training uses 4,965 identity-grouped examples; evaluation is the held-out fold ($n = 1{,}279$). Targets are constrained `<verdict>` tags. The student is a LoRA adapter on Qwen2.5-7B-Instruct. This split is the v052mix checker suite, not the 1,472-item program-logic panel.

Table 8: Held-out fold-0 agreement with the `gpt-oss-120b` teacher ($n = 1{,}279$). Acc. includes unparsed outputs. FR/FA are rates on teacher-correct / teacher-incorrect items.

| Model | Acc. (incl. unclear) | Unclear | FR | FA |
|---|---|---|---|---|
| 7B Instruct base | 0.00 | 99.92 | 0.18 | 0.00 |
| 7B LoRA student | 99.06 | 0.00 | 0.54 | 1.24 |

## E Extra-Seed Same-Scale Preserve

Table 9 repeats the Table 4 7B joint-text preserve-first protocol on Dual Routing and Shared extra seeds of the same 1,472-item panel (Dual-45/47/48 use clean reruns). The guard still improves both variants and cuts false rejection, with the larger recovery on Shared.

Table 9: Extra Dual/Shared seeds 45–48, 1,472-item panel, 7B logistic preserve-first veto. Dual-45/47/48 are clean reruns. Values are percentages except $\Delta$ (pp).

| Variant | Current | Preserve | $\Delta$ | FR $0\to1$ |
|---|---|---|---|---|
| Dual read/write | 38.99 | 41.83 | $+2.84$ | $4.28\to0.17$ |
| Shared adapter | 32.68 | 47.61 | $+14.93$ | $18.09\to0.10$ |

## F 7B NEURAL PRESERVE ON EXTRA DUAL/SHARED SEEDS

The Appendix D student can be applied as a preserve-first veto without a logistic on solver/checker text: a parsed `<verdict>correct</verdict>` keeps the solver, any other parsed or unparsed output leaves the current accept/revise action unchanged, and a revision is never added. Table 10 applies this overlay to extra Dual/Shared seeds 45–48 of the 1,472-item program-logic panel (Dual-45/47/48 are clean reruns). Extra Shared mean final accuracy rises from $32.68\%$ to $45.67\%$ ($+12.99$ pp); extra Dual rises only from $38.99\%$ to $39.71\%$ ($+0.71$ pp). Unclear rate is $0$ on every seed. This is a retrospective overlay, not an online jointly trained controller, and it is not a Dual-versus-Shared headline.

Table 10: Extra Dual/Shared seeds 45–48, 1,472-item panel, 7B LoRA student as preserve-first veto. Dual-45/47/48 are clean reruns. Values are percentages except $\Delta$ (pp) and unclear (rate).

| Panel | Current | Preserve | $\Delta$ | FR $0\to1$ | Unclear |
|---|---|---|---|---|---|
| Shared 45 | 41.10 | 51.29 | $+10.19$ | $13.99\to2.38$ | 0.00 |
| Shared 46 | 35.80 | 45.52 | $+9.71$ | $16.24\to2.79$ | 0.00 |
| Shared 47 | 25.82 | 40.76 | $+14.95$ | $21.54\to1.83$ | 0.00 |
| Shared 48 | 27.99 | 45.11 | $+17.12$ | $20.58\to2.38$ | 0.00 |
| Shared mean | 32.68 | 45.67 | $+12.99$ | $18.09\to2.34$ | 0.00 |
| Dual 45 rerun | 36.28 | 36.68 | $+0.41$ | $4.76\to0.75$ | 0.00 |
| Dual 46 | 41.58 | 42.53 | $+0.95$ | $4.42\to1.49$ | 0.00 |
| Dual 47 rerun | 40.76 | 41.78 | $+1.02$ | $3.94\to0.61$ | 0.00 |
| Dual 48 rerun | 37.36 | 37.84 | $+0.48$ | $4.01\to0.82$ | 0.00 |
| Dual mean | 38.99 | 39.71 | $+0.71$ | $4.28\to0.92$ | 0.00 |

Table 11 repeats the overlay on the frozen evaluation logs behind Table 4. Shared and the three collapsing controls all improve. Dual Routing is mixed: seed 43 gains $+1.77$ pp, but seeds 42 and 44 lose $2.11$ and $2.85$ pp, so the three-seed Dual pool ($-1.06$ pp) is not folded as a Dual win. Those two held seeds are exactly the Dual replicates whose current gate already had positive system gain ($+8.22$ and $+5.16$); a parsed `correct` blocks beneficial revisions there. Across all 23 extra and main overlays, Spearman correlation between current gain and neural $\Delta$ is $-0.99$ (18 negative-gain panels, mean $\Delta = +16.72$; 5 non-negative-gain panels, mean $\Delta = -0.62$).

Table 11: Primary seeds 42–44, 1,472-item panel, 7B LoRA student as preserve-first veto. Dual 42/44 are held ($\Delta \leq 0$). Values are percentages except $\Delta$ (pp). Unclear rate is $0$ on every seed.

| Panel | Current | Preserve | $\Delta$ | FR $0\to1$ |
|---|---|---|---|---|
| Shared mean | 26.92 | 44.32 | $+17.39$ | $22.44\to2.92$ |
| Role-isolated mean | 7.84 | 23.98 | $+16.15$ | $21.13\to4.60$ |
| Projected-gradient A mean | 4.94 | 25.50 | $+20.56$ | $26.34\to4.46$ |
| Projected-gradient B mean | 0.34 | 28.28 | $+27.94$ | $34.58\to6.61$ |
| Dual 42 (held) | 21.60 | 19.50 | $-2.11$ | $1.49\to0.48$ |
| Dual 43 | 37.50 | 39.27 | $+1.77$ | $4.69\to1.22$ |
| Dual 44 (held) | 26.15 | 23.30 | $-2.85$ | $2.38\to0.34$ |

A pad-free canonical-tag log-likelihood ratio, with $\tau$ chosen to maximize extra Shared $\Delta$ subject to every extra Dual seed $\Delta \geq 0$, still yields Dual-42/44 losses ($-3.80$ and $-3.26$). The only $\tau$ that keeps both of those seeds non-negative never preserves and returns $\Delta = 0$. We therefore keep the parsed-tag student and do not fold a Dual-safe LLR overlay.

## G UTILITY PRESERVE STUDENT ON HELD-OUT SEEDS

The parsed-tag student judges solver correctness. A preserve-first controller instead needs the sign of revision utility. We train a rank-16 Qwen2.5-7B LoRA to emit `<decision>preserve</decision>` unless the revision strictly beats the solver, and `revise` if it does. Training uses only extra Dual/Shared seeds 45–48, with labels balanced to equal counts. Table 12 is a five-fold problem-text holdout: each test problem is absent from that fold's training file, and the five slices cover the primary seeds 42–44. At test time a `preserve` decision keeps the solver and a `revise` decision leaves the current action unchanged, so the policy never adds a revision. Unclear rate is $0$.

Under that holdout, an argmax decision raises Shared from $26.92\%$ to $48.23\%$ ($+21.31$ pp), above both the parsed-tag student ($+17.39$) and the in-panel logistic ($+19.86$). A same-problem seed transfer gives $+21.38$ pp, so the gain is not memorization of the prompt. That argmax point still loses on Dual-42 ($-1.97$). Table 13 cross-fits the threshold: for each problem fold the threshold maximizes Shared $\Delta$ on the other folds subject to Dual-42/43/44 on those folds having $\Delta \geq 0$, then scores only the held-out fold. One 588-item Dual-42 split was $+0.85$, but it does not survive this cross-fit. Full-panel Dual-42 is $-0.07$ (item bootstrap $95\%$ interval $[-0.82, +0.68]$, which includes zero). Shared remains $+20.67$ $[+19.45, +21.97]$, still above the in-panel logistic ($+19.86$). Dual-43 is $+2.72$ $[+1.56, +3.87]$.

Table 12: Utility preserve student, five-fold problem-text holdout, trained on extra seeds 45–48 and tested on primary seeds 42–44. Dual-42 is held. Values are percentages except $\Delta$ (pp).

| Panel | Current | Preserve | $\Delta$ | FR 0→1 |
|---|---|---|---|---|
| Shared mean | 26.92 | 48.23 | +21.31 | 22.44→0.14 |
| Role-isolated mean | 7.84 | 28.69 | +20.86 | 21.13→0.16 |
| Projected-gradient A mean | 4.94 | 30.12 | +25.18 | 26.34→0.82 |
| Projected-gradient B mean | 0.34 | 32.61 | +32.27 | 34.58→2.29 |
| Dual 42 (held) | 21.60 | 19.63 | −1.97 | 1.49→0.34 |
| Dual 43 | 37.50 | 39.95 | +2.45 | 4.69→0.14 |
| Dual 44 | 26.15 | 26.36 | +0.20 | 2.38→0.20 |

Table 13: Cross-fit preserve threshold on the identity-held-out utility student. Every item is scored with a threshold fit on the other problem folds. Dual-42 is the full 1,472-item seed and is not a win. Values are percentages except $\Delta$ (pp).

| Panel | Current | Preserve | $\Delta$ | FR 0→1 |
|---|---|---|---|---|
| Shared mean | 26.92 | 47.60 | +20.67 | 22.44→1.15 |
| Role-isolated mean | 7.84 | 27.42 | +19.59 | 21.13→1.47 |
| Projected-gradient A mean | 4.94 | 28.74 | +23.80 | 26.34→2.38 |
| Projected-gradient B mean | 0.34 | 29.60 | +29.26 | 34.58→5.30 |
| Dual 42 (full seed) | 21.60 | 21.54 | −0.07 | 1.49→0.54 |
| Dual 43 | 37.50 | 40.22 | +2.72 | 4.69→0.61 |
| Dual 44 | 26.15 | 26.97 | +0.82 | 2.38→0.41 |

Table 14 replaces the utility labels with only the decisive cases (solver right and redo wrong, or the reverse) and continues each identity-fold adapter for one epoch. The threshold is cross-fit: each fold maximizes Shared $\Delta$ on the other folds subject to that Dual-42 slice staying at $\Delta \geq -0.07$. Primary Shared mean $\Delta = +21.4221$ and full-panel Dual-42 is $+0.2717$. The earlier cross-fit, Dual-42 at $-0.07$, remains the untuned reference and is still a wash relative to the $+0.5$ bar.

Table 14: Decisive-only utility continuation with a cross-fit preserve threshold. Dual-42 is the full 1,472-item seed and is not a $+0.5$ win. Values are percentages except $\Delta$ (pp).

| Panel | Current | Preserve | $\Delta$ |
|---|---|---|---|
| Shared 42 | 16.44 | 46.88 | +30.4348 |
| Shared 43 | 22.96 | 52.85 | +29.8913 |
| Shared 44 | 41.37 | 45.31 | +3.9402 |
| Shared mean | 26.92 | 48.35 | +21.4221 |
| Dual 42 | 21.60 | 21.88 | +0.2717 |
| Dual 43 | 37.50 | 40.63 | +3.1250 |
| Dual 44 | 26.15 | 27.11 | +0.9511 |

The reverse split trains on primary seeds 42–44 and tests extra seeds 45–48 on held-out problems ($n = 1472$ each, unclear rate 0). Extra Shared mean $\Delta = +14.7758$ and extra Dual mean $\Delta = +0.2548$, so both sides of the pre-registered bar are met. Extra Dual-48 is $-0.6114$; the mean stays non-negative because the other three extra Dual seeds are positive. This does not change the Dual-42 cross-fit.

Table 17 cross-fits $\log p(\text{preserve}) - \log p(\text{revise}) > \tau$. For each problem fold, $\tau$ maximizes extra Shared $\Delta$ on the other folds subject to every extra Dual seed on those folds having $\Delta \geq 0$. The first such cross-fit has extra Shared mean $\Delta = +15.0985$ and Dual-48 at $+0.1359$. The decisive-only

Table 15: Same decisive-only adapters. Each fold's threshold is the one that maximizes Shared $\Delta$ on the other folds among thresholds with Dual-42 $\Delta \geq +0.5$ and Shared $\Delta \geq +18$ on those folds. Held-out Dual-42 is $+0.6114$. Role-isolated and both projected-gradient rows use those same frozen thresholds and were not used to choose them.

| Panel | Current | Preserve | $\Delta$ |
|---|---|---|---|
| Shared mean | 26.92 | 48.01 | $+21.0824$ |
| Role-isolated mean | 7.84 | 27.76 | $+19.9275$ |
| Projected-gradient A mean | 4.94 | 28.03 | $+23.0978$ |
| Projected-gradient B mean | 0.34 | 29.89 | $+29.5516$ |
| Dual 42 | 21.60 | 22.21 | $+0.6114$ |
| Dual 43 | 37.50 | 40.56 | $+3.0571$ |
| Dual 44 | 26.15 | 27.38 | $+1.2228$ |

Table 16: Utility preserve student trained on primary seeds 42–44 and tested on extra seeds 45–48, problem text held out. Values are percentages except $\Delta$ (pp).

| Panel | Current | Preserve | $\Delta$ |
|---|---|---|---|
| Shared 45 | 41.10 | 52.04 | $+10.94$ |
| Shared 46 | 35.80 | 47.01 | $+11.21$ |
| Shared 47 | 25.82 | 44.43 | $+18.61$ |
| Shared 48 | 27.99 | 46.33 | $+18.34$ |
| Shared mean | 32.68 | 47.45 | $+14.7758$ |
| Dual 45 rerun | 36.28 | 36.75 | $+0.48$ |
| Dual 46 | 41.58 | 42.26 | $+0.68$ |
| Dual 47 rerun | 40.76 | 41.24 | $+0.48$ |
| Dual 48 rerun | 37.36 | 36.75 | $-0.61$ |
| Dual mean | 38.99 | 39.25 | $+0.2548$ |

continuation in Table 17 raises extra Shared to $+15.7948$ and Dual-48 to $+1.0190$. The full-panel Dual-42 cross-fit on the primary seeds remains a separate result.

Table 17: Cross-fit LLR threshold after a decisive-only continuation of the utility student trained on primary seeds and tested on extra seeds. Every item uses a threshold fit on the other folds. Values are percentages except $\Delta$ (pp).

| Panel | Current | Preserve | $\Delta$ |
|---|---|---|---|
| Shared 45 | 41.10 | 52.58 | $+11.4810$ |
| Shared 46 | 35.80 | 49.39 | $+13.5870$ |
| Shared 47 | 25.82 | 45.58 | $+19.7690$ |
| Shared 48 | 27.99 | 46.33 | $+18.3424$ |
| Shared mean | 32.68 | 48.47 | $+15.7948$ |
| Dual 45 rerun | 36.28 | 39.20 | $+2.9212$ |
| Dual 46 | 41.58 | 44.09 | $+2.5136$ |
| Dual 47 rerun | 40.76 | 43.00 | $+2.2418$ |
| Dual 48 rerun | 37.36 | 38.38 | $+1.0190$ |

### G.1 Does the Preserve Rule Need Its Own Threshold?

The cross-fit above holds out *problems*: $\tau$ for fold $k$ is fit on the other folds, so no item scores under a threshold that saw it. That makes the item-level claim honest but leaves a seed-level gap. The seeds whose non-negativity the constraint enforces are exactly the four extra Dual seeds the table reports, so $\tau$ was chosen already knowing which checkpoints had to come out non-negative. That design therefore cannot say whether the rule helps a training run it was not tuned on.

Two facts make the gap addressable without new training. First, the utility adapter is trained on primary seeds 42–44 and every panel here is an extra seed 45–48, so the adapter is already seed-held-out and $\tau$ was the sole remaining path from a scored checkpoint to its own rule. Second, folds are keyed on problem text, so fold $k$ is the same problem set in every panel and a panel can be removed from the selection set while the fold structure survives. Table 18 adds that second holdout dimension. *Panel-held* drops the scored panel from both the objective and the constraint; *seed-held* drops both panels sharing its seed index. The objective, the constraint and the grid are otherwise those of Table 17.

Table 18: Extra-seed preserve $\Delta$ (pp) under progressively stricter holdouts, after the decisive-only continuation. *Argmax* fits no threshold and obeys the emitted decision; the three cross-fits differ only in what is removed from the $\tau$ selection set. Brackets are $10^4$-resample item bootstrap intervals. $n = 1472$ per panel, unclear rate 0.

| | | Argmax (no $\tau$) | | Cross-fit $\tau$ | | |
|---|---|---|---|---|---|---|
| Panel | Current | Before | After cont. | Problem | Panel-held | Seed-held |
| Shared 45 | 41.10 | +10.94 | +11.41 [+9.44, +13.45] | +11.48 | +10.80 | +11.55 |
| Shared 46 | 35.80 | +11.21 | +13.38 [+11.28, +15.56] | +13.59 | +13.45 | +13.59 |
| Shared 47 | 25.82 | +18.61 | +19.70 [+17.46, +21.94] | +19.77 | +19.84 | +19.77 |
| Shared 48 | 27.99 | +18.34 | +18.34 [+16.10, +20.58] | +18.34 | +17.19 | +17.19 |
| Dual 45 rerun | 36.28 | +0.48 | +2.79 [+1.43, +4.14] | +2.92 | +3.06 | +2.92 |
| Dual 46 | 41.58 | +0.68 | +2.51 [+1.15, +3.74] | +2.51 | +2.51 | +2.51 |
| Dual 47 rerun | 40.76 | +0.48 | +1.97 [+0.68, +3.12] | +2.24 | +2.31 | +2.45 |
| Dual 48 rerun | 37.36 | −0.61 | +0.68 [−0.82, +1.97] | +1.02 | +0.88 | +2.17 |
| Shared mean | — | +14.78 | +15.71 | +15.79 | +15.32 | +15.52 |
| Dual mean | — | +0.25 | +1.99 | +2.17 | +2.19 | +2.51 |
| Panels $> 0$ | — | 7/8 | 8/8 | 8/8 | 8/8 | 8/8 |

The strongest column is the one that fits nothing. Obeying the student's emitted decision after the decisive-only continuation makes all eight panels positive, extra Shared mean $+15.7099$ and extra Dual mean $+1.9871$, with the weakest panel Dual-48 at $+0.6793$. Before the continuation the same untuned rule left Dual-48 at $-0.6114$ (Table 18, *Before*); the continuation is what removes the one negative seed, and it does so on a rule with no free parameter. This clears the pre-registered extra-seed bar without a threshold, so for this operating point the tuning question does not arise. The gain is not uniform: Shared-48 is unchanged to four decimals while Dual-45 moves $+2.31$.

Tightening the holdout does not remove the effect. All 32 cells across the four designs are positive, and both arm means stay positive in every design. Panel-held costs extra Shared $0.48$ pp relative to the cross-fit of Table 17 and Dual-48 $0.14$ pp. Seed-held reports a *larger* Dual mean than the looser designs, which should not be read as the stricter protocol being kinder. Dual-48 prefers a large threshold, its own oracle optimum is 7.75, while Shared-45/47/48 prefer small ones; removing seed 48 from the selection set leaves an objective that happens to settle near $\tau \approx 5.75$, which suits Dual-48. That is a coincidence of the selection objective, not evidence of a stronger rule.

A threshold sweep makes the point directly. Over the full LLR support ($\tau \in [-17, 18]$, step $0.125$, 281 points, 11,553 distinct LLR values), all eight panels are strictly positive on a single contiguous interval $\tau \in [-2.125, +15.000]$, width 17.125 and 138 of 281 grid points. That interval contains the untuned $\tau = 0$, and the worst single panel anywhere inside it is $+0.0679$. The result is therefore a property of the student rather than of a tuned operating point. Four guards are recorded with the artifact: the Table 17 cross-fit reproduces to $10^{-6}$ on all eight panels and all five thresholds, every pooled panel is exactly 1,472 unique items, no panel appears in its own selection set, and the generation-text scorer agrees with $\mathrm{llr} > 0$ to within one item per panel (16 disagreements in 11,776 items). Still, $n = 4$ replicates per arm is descriptive; we claim no seed-level significance, and none of these numbers is an extra-seed final-accuracy headline or a solution to Dual-42. The supplementary script `replay_preserve_holdouts.py` recomputes every cell of Table 18 and the sweep from the per-item decisions (Appendix J).

### G.2 Transfer to Checkpoints Created After the Rule Was Frozen

Every panel in Table 18 existed when the decisive-only continuation was built. As a stricter test we took those five fold adapters unchanged and, before scoring anything, preregistered their application to twelve checkpoints trained afterwards: role-isolated, projected-gradient A, and projected-gradient B at seeds 45–48, all evaluated on the same 1,472-item panel at step 100. The adapters were trained on seeds 42–44 of all five training variants, so these are new training runs of variants the student has seen, scored on problems it never trained on. The rule is unchanged: obey the emitted decision with no threshold, each item scored by the adapter of its own problem fold, and every panel reported whatever its sign.

All twelve panels are positive (Table 19); the smallest gain is $+8.42$ pp (Role-isolated 45) and the arm means are $+15.54$, $+31.67$, and $+32.80$ pp. One of 17,664 generations carried no decision tag and kept its current action. The gains are large because these workflows revise almost every answer and so destroy many correct ones: mean false rejection is $28.18\%$ before the rule and $0.21\%$ after. The rule restores $4{,}941$ of the $4{,}978$ harmful revisions, but it also vetoes 230 of the 348 beneficial ones. That cost is concentrated where revision still helps: Role-isolated 45 recovers $44.6\%$ of its preserve-only ceiling, and on nine of the twelve panels the rule reaches at least $98\%$. The trivial rule that never revises is nearly as good on these panels (mean $0.46$ pp behind the student; the student is ahead on 6 of 12). The result therefore says the frozen rule transfers to new training runs whose revisions mostly hurt, by recovering what never revising would; it does not say it separates good revisions from bad ones on a workflow that already revises well. That is what the Dual rows of Table 18 test, and there the student beats both the current gate and never revising on all four Dual panels ($+0.68$ to $+2.79$ pp above the better of the two).

Table 19: Frozen preserve rule on twelve checkpoints trained after it was frozen, preregistered. $\Delta$ in pp with $10^4$-resample item bootstrap intervals; current and preserve are final-accuracy percentages; FR is false rejection. $n = 1472$ per panel.

| Panel | Current | Preserve | $\Delta$ | FR |
|---|---|---|---|---|
| Role-isolated 45 | 23.91 | 32.34 | $+8.42$ [$+5.77$, $+11.14$] | 18.89→0.14 |
| Role-isolated 46 | 0.34 | 27.11 | $+26.77$ [$+24.52$, $+29.01$] | 27.11→0.34 |
| Role-isolated 47 | 6.11 | 15.42 | $+9.31$ [$+7.40$, $+11.21$] | 12.50→0.68 |
| Role-isolated 48 | 0.20 | 17.87 | $+17.66$ [$+15.69$, $+19.63$] | 18.00→0.27 |
| Projected-gradient A 45 | 1.09 | 36.75 | $+35.67$ [$+33.15$, $+38.04$] | 35.94→0.27 |
| Projected-gradient A 46 | 1.56 | 30.37 | $+28.80$ [$+26.49$, $+31.11$] | 29.21→0.41 |
| Projected-gradient A 47 | 12.30 | 28.33 | $+16.03$ [$+13.79$, $+18.27$] | 18.68→0.07 |
| Projected-gradient A 48 | 1.70 | 47.89 | $+46.20$ [$+43.68$, $+48.78$] | 46.26→0.00 |
| Projected-gradient B 45 | 1.83 | 34.78 | $+32.95$ [$+30.57$, $+35.39$] | 33.08→0.07 |
| Projected-gradient B 46 | 7.40 | 41.98 | $+34.58$ [$+32.20$, $+37.02$] | 34.65→0.07 |
| Projected-gradient B 47 | 0.41 | 32.68 | $+32.27$ [$+29.89$, $+34.65$] | 32.34→0.07 |
| Projected-gradient B 48 | 1.97 | 33.36 | $+31.39$ [$+29.01$, $+33.83$] | 31.52→0.14 |

### G.3 Consequence Labels versus Correctness Labels at a Matched Veto Budget

A preserve rule can be driven by either kind of 7B student: the correctness student of Appendix D, which estimates whether the solver answer is right, or the decisive-only consequence student, which estimates whether the revision beats it. Both score every item of the eight extra-seed panels without having trained on it. To compare their rankings rather than their thresholds, we fix the number of revisions each may veto on a panel at the number the consequence student vetoes under its untuned argmax, and let each student veto its most preserve-leaning revisions. This matched-budget comparison is a post hoc diagnostic, not preregistered (Table 20).

Table 20: Final accuracy (percent) when a correctness-trained and a consequence-trained 7B student veto the same number of the gate's revisions on each extra-seed panel. Budget is vetoes over the gate's revisions.

| Panel | Gate | Never revise | Budget | Correctness | Consequence | $\Delta$ |
|---|---|---|---|---|---|---|
| Dual 45 | 36.28 | 36.21 | 439/537 | 36.55 | 39.06 | $+2.51$ |
| Dual 46 | 41.58 | 41.51 | 531/612 | 41.92 | 44.02 | $+2.11$ |
| Dual 47 | 40.76 | 40.90 | 542/612 | 41.03 | 42.66 | $+1.63$ |
| Dual 48 | 37.36 | 36.68 | 601/633 | 36.75 | 37.98 | $+1.22$ |
| Shared 45 | 41.10 | 51.63 | 1065/1106 | 51.90 | 52.51 | $+0.61$ |
| Shared 46 | 35.80 | 46.67 | 1227/1302 | 46.54 | 49.18 | $+2.65$ |
| Shared 47 | 25.82 | 41.24 | 1193/1411 | 41.30 | 45.52 | $+4.21$ |
| Shared 48 | 27.99 | 46.40 | 1446/1462 | 46.40 | 46.33 | $-0.07$ |

At the same veto budget, the consequence student is ahead on seven of eight panels, by $+1.86$ points on average (panel-level $t$ interval $[+0.75, +2.97]$). The two students differ in training data as well as label; the preregistered same-data comparison follows.

**Same-data preserve-first students (preregistered).** We trained a twin of the one-stage student that differs only in its label: the same rows (seeds 42–44, all five training variants, problems outside the held-out fold), prompt, class balancing, adapter, and schedule, with `preserve` meaning "the solver answer is correct". All five problem folds completed; each item is scored by the adapter of its own fold under the untuned argmax (Table 21).

Table 21: Same-data 7B preserve-first students on held-out problems, all five folds (1,472 items per panel), preregistered. Final accuracy in percent, means over panels.

| Panels | Gate | Never revise | Consequence label | Correctness label | Frozen decisive |
|---|---|---|---|---|---|
| Dual (4) | 38.99 | 38.82 | 39.25 | **42.80** | 40.98 |
| Shared (4) | 32.68 | 46.48 | 47.45 | **49.71** | 48.39 |
| Post-freeze (12) | 4.90 | 31.11 | 31.25 | 31.47 | **31.57** |

Every preserve-first student improves on the deployed gate. The correctness-label student is the strongest on the eight extra-seed panels: it beats both the deployed gate and never revising on all four Dual panels ($+3.80$ and $+3.97$ points on average) and all four Shared panels ($+17.03$ and $+3.23$), and it improves every one of the 20 panels over its gate. On the problem-disjoint split below, it improves all 35 workflows over their gates (mean $+4.52$ points) and beats both the gate and never revising on all seven Dual workflows. The preserve-first rule is thus robust to how its 7B student is supervised.

### G.4 A Problem-Disjoint Split with Six Unseen Task Families

Every earlier test used the 1,472-item program panel. We therefore preregistered, before scoring, the same five frozen adapters on a split none of them had seen: the 184-item validation mix of the concurrent routing study, which shares no question with that panel and adds APPS, ZebraLogic, 24-game, maze, sudoku, and ordering problems to the four program templates. It covers all 35 routing checkpoints (five training variants, seeds 42–48) at step 100, 6,440 workflow–item records. Because no problem belongs to a training fold, the five adapters vote and a majority for `preserve` keeps the solver answer; nothing is retuned, and every workflow is reported (Table 22).

Table 22: Frozen preserve student on the problem-disjoint 184-item split, preregistered. Final accuracy (percent), mean over seven checkpoints per training variant; $\Delta$ is student minus the deployed gate (pp), with the number of positive checkpoints.

| Variant | Gate | Never revise | Always revise | Student | $\Delta$ | Pos. |
|---|---|---|---|---|---|---|
| Shared | 10.64 | 14.83 | 10.33 | 14.98 | $+4.35$ | 6/7 |
| Role-isolated | 4.19 | 8.54 | 3.03 | 8.93 | $+4.74$ | 7/7 |
| Projected-gradient A | 4.50 | 11.34 | 4.27 | 11.65 | $+7.14$ | 7/7 |
| Projected-gradient B | 2.33 | 10.64 | 1.86 | 10.56 | $+8.23$ | 7/7 |
| Dual read/write | 12.19 | 10.48 | 13.28 | 11.57 | $-0.62$ | 2/7 |

The rule transfers where it should and fails where it should. It raises final accuracy on 29 of 35 workflows (mean $+4.77$ pp; no parse failures) and on APPS ($+6.19$ pp) and ZebraLogic ($+0.45$), families it never trained on. It distinguishes harmful from beneficial revisions: of the gates' 407 harmful revisions it lets 4 through ($1.0\%$), of their 124 beneficial ones 28 ($22.6\%$), and it beats a veto of the same number of revisions chosen at random on 34 of 35 workflows. But it is conservative. Averaged over workflows it is only $0.37$ pp above never revising, and on the Dual workflows, whose revisions already help (always revise $13.28\%$ versus never $10.48\%$), it loses to the deployed gate on four of seven. On new problems the frozen rule is therefore a reliable guard against destructive revision rather than a way to keep the repairs of a gate that already works.

## H Pipeline and Fixed-Candidate Controls

False rejection under specialized versus shared checking is heterogeneous: lower, equal, and higher in two, one, and three cells, respectively. Endpoint accuracy rises in all six cells, so we do not describe the specialization effect as a uniform false-rejection reduction.

Table 23: Fixed first-pass solver and reviser adapter; checker feedback and the conditioned revision may change. Final accuracy of the specialized and shared pipelines in percent; gains of the specialized pipeline in percentage points.

| Panel | Seed | Specialized | Shared | vs. shared verifier | vs. solver | vs. always revise |
|---|---|---|---|---|---|---|
| Sparse-clue | 154 | 15.23 | 10.81 | +4.427 | +8.724 | +1.172 |
| Sparse-clue | 155 | 16.02 | 11.20 | +4.818 | +9.375 | +0.911 |
| Strict-no-position | 154 | 16.15 | 12.63 | +3.516 | +9.896 | +1.302 |
| Strict-no-position | 155 | 16.80 | 11.07 | +5.729 | +10.677 | +1.302 |
| Constraint-stress | 154 | 17.12 | 13.48 | +3.646 | +11.458 | +1.302 |
| Constraint-stress | 155 | 17.71 | 13.28 | +4.427 | +12.305 | +1.042 |

Table 24: Specialized-minus-shared pipeline effects with structural-family bootstrap intervals. The first answer and reviser adapter are fixed; feedback- conditioned revision text may differ.

| Panel | Replicate | Effect (pp) | 95% interval (pp) |
|---|---|---|---|
| Sparse-clue | 1 | +4.427 | [+1.733, +7.231] |
| Sparse-clue | 2 | +4.818 | [+2.152, +7.451] |
| Strict-no-position | 1 | +3.516 | [+0.980, +5.874] |
| Strict-no-position | 2 | +5.729 | [+2.592, +8.549] |
| Constraint-stress | 1 | +3.646 | [+1.170, +6.285] |
| Constraint-stress | 2 | +4.427 | [+2.074, +6.715] |

# I Auxiliary Baseline Boundaries

The threshold diagnostic and PAG profiles use different protocols and sample counts; Table 26 therefore reports them in separate blocks. The PAG sign reverses under a stronger domain-neutral prompt profile, which rules out a profile-independent superiority claim.

The released-profile comparison uses the same Qwen2.5-7B base, 80 rows over 69 identities, ten updates, and the same sparse-clue questions. It is not matched for total supervision or wall-clock compute. We retain it only as an auxiliary boundary rather than primary evidence.

# J Anonymous Reproduction Contract

The accompanying archive contains a JSON evidence ledger and a Python standard-library replay. Their SHA-256 digests begin `84d3c5d2eec2` and `d05b90976883`; the manifest records the full values. Running `reproduce_tables.py` checks retained aggregate summaries for the 18-run association, 114-case controller comparison, 458-row accounting including the nested cross-fit, six pipeline cells, six matched-random controls, the fold-assignment digest, and compute arithmetic. It terminates with `All anonymous evidence checks passed.` The archive also contains the per-item preserve decisions for the eight extra-seed panels of Appendix G.1 and the twelve post-freeze panels of Appendix G.2 (SHA-256 prefix `2b8086875071`), with each item's fold, outcomes, and log-likelihood ratio. `replay_preserve_holdouts.py` (`fffdb636f51d`) recomputes every cell of Tables 18 and 19 and the threshold sweep from that file alone and terminates with `All preserve-holdout checks passed.` A third script (`63151f7aa7ce`) recomputes the episode-clustered intervals of Table 2 and the results of Tables 22 and 20 from their shipped per-item records.

Table 25: Same-payload, family-revision-count-matched randomization. Candidates, checker feedback, and model-call counts are fixed; response-token counts vary with the selected revision lengths.

| Panel | Replicate | Revision rate (%) | Gain over random (pp) | Exact tail probability |
|---|---|---|---|---|
| Sparse-clue | 1 | 66.67 | +2.558 | $7.00 \times 10^{-8}$ |
| Sparse-clue | 2 | 67.58 | +2.602 | $1.86 \times 10^{-7}$ |
| Strict-no-position | 1 | 63.28 | +3.210 | $2.79 \times 10^{-10}$ |
| Strict-no-position | 2 | 65.76 | +3.261 | $1.05 \times 10^{-10}$ |
| Constraint-stress | 1 | 62.70 | +4.160 | $1.88 \times 10^{-27}$ |
| Constraint-stress | 2 | 63.54 | +4.168 | $1.88 \times 10^{-26}$ |

Table 26: Auxiliary diagnostics, reported with their protocol boundaries.

| Method/profile | Accuracy (%) | Budget |
|---|---|---|
| Threshold ($n = 1{,}578$) | 29.214 | 73.384% revised |
| KEEPER ($n = 1{,}578$) | 29.214 | 60.710% revised |
| Threshold at 60.710% | 27.693 | 60.710% revised |
| Joint score at 60.710% | 28.010 | 60.710% revised |
| Released PAG profile ($n = 768$) | 10.417 | 100% tokens |
| KEEPER ($n = 768$) | 14.974 | 58.23% tokens |
| Domain-neutral PAG ($n = 768$) | 19.531 | Profile diagnostic |