CONSENSUS IS NOT CORRECTNESS: RELIABILITY-AWARE SELF-TEACHING FOR UNSUPERVISED SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised on-policy self-distillation uses a model's own consensus to construct training signal, yet answer agreement does not certify the reasoning trace being imitated. We ask which quantities consensus actually certifies and introduce Reliability-Aware Self-Teaching (RAST), a control layer that separates answer confidence, group trace consistency, teacher-process reliability, and retained supervision. In process-auditable diagnosis logs, answer-correct traces can fail process audit in both Gemini 2.5 Pro and a dedicated Qwen3-4B process-gold audit subset, although process measurability is selective and the conditional rates are not population prevalence estimates. Answer confidence and teacher reliability specialize to different audit targets (AUROC 0.894 vs. 0.624 and 0.649 vs. 0.788), while joint thresholding traces a risk–coverage frontier below the audit-defined risk rate expected under random thinning of the same candidate pool. Correction-time selector rankings also disagree between answer and process objectives. A frozen five-seed LoRA study does not establish a multiplicity-corrected accuracy winner. A post-freeze three-seed MATH-500 downstream stress test likewise does not resolve a Majority@8 advantage for either self-distillation condition: the RAST–U-OPSD difference is -0.07 pp (95% paired two-way bootstrap CI $[-1.40, +1.33]$). The supported conclusion is narrower: self-distillation should monitor answer trust, process trust, and supervision coverage separately rather than collapse them into a single consensus score.

1 INTRODUCTION

Self-distillation promises an attractive form of self-improvement: a language model can sample its own trajectories and turn selected behavior into training signal without an external teacher. Recent U-OPSD makes this especially direct: multiple rollouts form a majority-consistent pseudo-solution; a teacher distribution is conditioned on a short pseudo-solution and distilled into prefixes of an incorrect on-policy completion (Li et al., 2026). The recipe exploits a real property of modern models: majority aggregation can recover latent competence that a single rollout expresses unreliably (Wang et al., 2023a).

The same recipe contains a structural mismatch. Teacher *selection* is largely decided by answer-level agreement, while the object copied by self-distillation is a full reasoning trace. A trace can be *Lucky Correct*—correct in its final answer but wrong in audited intermediates—and multiple traces can repeat the same erroneous derivation. Consensus therefore answers the question "what outcome recurs?" but not "which reasoning trace is safe or useful to imitate?" This distinction connects self-distillation to process supervision and chain-of-thought faithfulness: final-answer correctness alone does not certify intermediate reasoning (Lightman et al., 2023; Wang et al., 2023b; Turpin et al., 2023; Lanham et al., 2023).

We study this gap as an empirical control problem rather than assuming that every process error is causally harmful. RAST (Reliability-Aware Self-Teaching) is a reliability-control layer for self-distillation. It separates answer confidence C_A, group trace consistency C_S, and per-candidate teacher reliability R_T; optionally uses independent derivation families and problem-internal verification; and maps these signals to full-trace, reduced-trace, answer-only, down-weighted, or abstaining supervision.

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.

The name *Self-Teaching* refers to this control layer; the training setting remains unsupervised self-distillation.

Our evidence has four layers. First, process audits test whether answer-level and trace-level trust actually coincide, including an explicit bridge from a commercial API-model audit to a dedicated Qwen3-4B process-gold audit subset. Second, correction-time interventions test selector objectives and randomize process-correct versus Lucky-Correct teachers. Third, a frozen five-seed LoRA study tests whether selector/control variants change held-out behavior after weight updates. Fourth, a post-freeze MATH-500 downstream stress test stress-tests downstream transfer on a public mathematical-reasoning benchmark. The combined result is not a new state-of-the-art rule; it is an empirical decomposition of what consensus can and cannot certify.

Contributions.

- We show, in Gemini 2.5 Pro and an explicitly identified Qwen3-4B process-gold audit subset, that answer-correct sampled traces can fail an auditable process rubric; we treat the rates as conditional diagnostics rather than population prevalence.
- We separate pseudo-label correctness, teacher-process audit quality, and audit-defined update risk, showing with task-cluster uncertainty and threshold sensitivity that answer confidence and teacher reliability predict different targets and that joint thresholding yields a non-trivial risk–coverage frontier.
- Correction-time studies expose objective mismatch rather than a selector winner: answer-correction and process-correction rankings differ, and ceiling-limited randomized interventions do not identify a universal Lucky-Correct harm effect.
- In a frozen five-seed LoRA study (40/40 valid jobs), no intervention is multiplicity-corrected superior; the strongest point estimate therefore remains descriptive rather than a winner claim.
- In a post-freeze MATH-500 downstream stress test (500 tasks, three frozen seeds), neither the paper-structure-faithful U-OPSD condition nor RAST resolves a Majority@8 gain over the same-seed M0; RAST and U-OPSD differ by only -0.07 pp (95% CI $[-1.40, +1.33]$).

2 Related Work

Self-consistency and self-training. Self-consistency aggregates sampled reasoning paths to improve inference (Wang et al., 2023a). Self-generated reasoning, preferences, confidence, and rewards have subsequently been turned into training signals (Zelikman et al., 2022; Huang et al., 2023; Yuan et al., 2024; Prasad et al., 2025; Prabhudesai et al., 2025; Shafayat et al., 2025). These methods motivate label-free learning but do not by themselves establish that a consensus trace is a good teacher.

On-policy distillation. On-policy distillation learns from student-generated trajectories under a teacher distribution (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024). OPSD conditions a same-model teacher on privileged verified context, whereas U-OPSD constructs privileged context from the model's own majority-consistent generations (Zhao et al., 2026; Li et al., 2026). Recent work targets estimator fragility, teacher–student mismatch, and token-level uncertainty/reliability (Fu et al., 2026; Wang et al., 2026; Ke et al., 2026; Liu et al., 2026; Tan and Hong, 2026). Table 1 locates our narrower distinction: RAST separates answer-level trust, selected trace-instance trust, and retained supervision coverage before deciding what to imitate; it is not presented as a higher-accuracy replacement for these methods.

Process supervision and faithfulness. Process reward models supervise intermediate steps rather than only outcomes; human-labeled PRM800K and automatically constructed Math-Shepherd are prominent examples (Lightman et al., 2023; Wang et al., 2023b). RAST does not replace a PRM: it targets settings where no learned process judge or human step labels are assumed, using internal structure and mechanically checkable claims when available. Separately, chain-of-thought explanations can be unfaithful to a model's actual decision process (Turpin et al., 2023; Lanham et al., 2023). Our "Lucky Correct" label is narrower and operational: it denotes answer-correct sampled traces whose auditable stated intermediates fail a process rubric, without claiming access to the model's latent causal reasoning.

Table 1: Closest on-policy/self-distillation reliability controls. The comparison is by *unit of control*, not by benchmark score.

Work	Primary reliability/control unit	Distinction from RAST
OPSD (Zhao et al., 2026)	Privileged verified solution context	Uses external privileged reasoning; RAST studies label-free teacher admission.
U-OPSD (Li et al., 2026)	Answer consensus; shortest agreeing pseudo-solution	Consensus constructs context; no separate audit of the selected trace instance.
Fu / Wang (Fu et al., 2026; Wang et al., 2026)	OPD estimator, teacher–student mismatch, signal regulation	Optimization/signal pathology rather than answer-vs.-selected-trace trust.
Ke / Liu (Ke et al., 2026; Liu et al., 2026)	Token entropy, position, branch viability	Token-level weighting; RAST adds answer/process/coverage separation at the trace-instance level.
SSOPD (Tan and Hong, 2026)	Correct–wrong intra-group contrast	Uses correctness partition within a sampled group; RAST focuses on no-judge trace admission and abstention/coverage.
GATES (Stein et al., 2026)	Tutor consensus as an admission gate for full trajectories	Consensus gating is shared; RAST separates answer trust, selected-trace trust, and retained coverage rather than treating consensus as the process-reliability certificate.

Metamorphic verification. Metamorphic testing checks expected invariances when a conventional oracle is unavailable (Chen et al., 1998; Segura et al., 2016). We adapt this principle to reasoning traces and combine it with deterministic recomputation of asserted intermediates.

Selective prediction and risk–coverage. The reject-option literature evaluates reliability by trading coverage for conditional risk rather than forcing a prediction everywhere (Geifman and El-Yaniv, 2017). We use the same reporting geometry for self-supervision admission: the quantity at risk is an audit-defined supervision event, not ordinary classification error. This framing is descriptive of selective control and does not turn the audit label into a causal downstream-harm claim.

3 Auditing Consensus and Trace Reliability

3.1 Lucky Correct appears in both the audit model and the student logs

The process-auditable diagnostic subset contains 29 unique diagnosis tasks. For the commercial API audit we use Gemini 2.5 Pro. The sanitized frozen export records the product/model family but not a stable vendor-side snapshot identifier for every archived call, nor a single complete per-call table from which every execution timestamp and sampling control for the 3,197-rollout aggregate can be reconstructed. We therefore report only metadata actually preserved by the audit, do not impute missing API settings from later local experiments, avoid claiming bitwise API reproducibility, and treat possible provider drift as a limitation. Across 3,197 Gemini rollouts, 912 (28.5%) contain enough structured intermediate quantities to score the process rubric. Among the 692 process-measurable rollouts whose final answer is correct, 381 pass the process audit and 311 do not; thus 44.9% are *Lucky Correct* conditional on being answer-correct and process-measurable.

For the student-model bridge, we use one dedicated frozen local process-gold round using the Qwen3-4B 4-bit MLX stack, 29 diagnosis tasks, $G = 8$, and a 2048-token generation limit. The round therefore contains 232 rollouts (29 tasks $\times$ 8). All 232 have auditable process quantities by construction because every sampled task carries process gold; among 96 answer-correct rollouts, 72 fail the process audit (75.0%). This 100% within-round measurability is a property of the deliberately selected audit task set, not model-wide coverage, and must not be compared to Gemini's 28.5% broad-audit coverage. The round is from the same student model family but is not the exact rollout stream used by the final 40 weight-update jobs. Table 2 is therefore a diagnostic model-family bridge, not an estimate of the final training pool.

The mechanism is visible in an information-gain diagnosis case. The correct option exceeds the runner-up by only 0.0077 bits. A 6–2 majority selects the correct option even though supporting traces mis-compute intermediate entropies and gains; a coherence-based teacher score is 0.73 while hidden process accuracy of the selected trace is 0.25. We do *not* report a prevalence number for common-mode error because the frozen artifact has no task-level label that separates a genuinely shared derivation from repeated but independently generated error.

Table 2: Lucky-Correct audit bridge. Rates are conditional on answer-correct, process-measurable traces. The Qwen row is one dedicated Qwen3-4B/MLX process-gold round; its 100% measurability is by audit-task construction, not model-wide coverage.

Model arm	Total	Meas.	Ans.-correct	Proc.-fail	Lucky rate
Gemini 2.5 Pro	3,197	912	692	311	44.9%
Qwen3-4B audit	232	232	96	72	75.0%

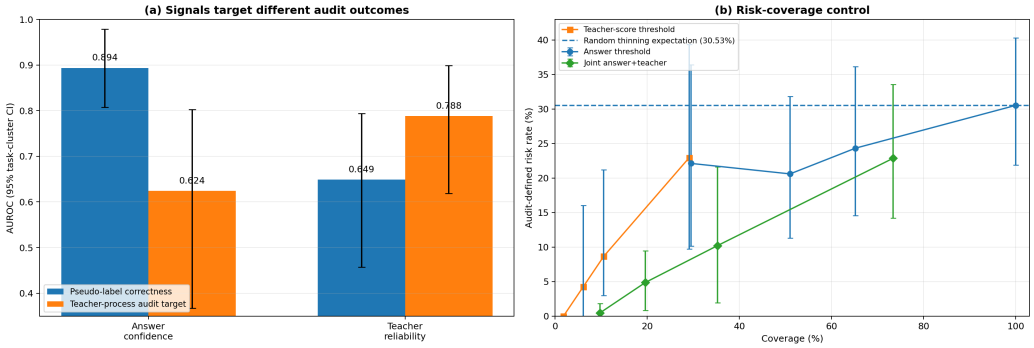

Figure 1: Diagnostic evidence with uncertainty. (a) Bars show AUROC; whiskers are 95% intervals from 10,000 task-cluster bootstrap replicates. (b) Frozen risk–coverage point estimates use the audit-defined update-risk label. Vertical whiskers are 95% task-cluster bootstrap sensitivity intervals. For the joint curve, interval widths come from an archived row-level reconstruction whose centers differ from the frozen summary by at most 0.13 percentage points in coverage and 0.96 points in risk; exact frozen points remain plotted. Random thinning independent of the audit label has expected risk 30.53% at every coverage (dashed line).

3.2 Answer trust, process trust, and audit-defined update risk are different targets

We define the audit targets before evaluating the signals. **Pseudo-label correctness** means that the majority key matches the hidden answer. The **teacher-process audit target** is positive when the selected teacher's decisive-quantity vector matches the frozen `process_gold` rubric on at least half of auditable quantities under the frozen relative-error rule; it is an automatic process-audit target, not a claim of causal downstream utility. We call the frozen-artifact `harmful_update` quadrant an **audit-risk event**: the consensus teacher/pseudo-label is incorrect while correct behavior exists in the sampled group, so imitation could displace a correct minority. Its admitted fraction is the **audit-defined update risk**. This is a diagnostic risk indicator, not evidence that an individual gradient update causally degrades downstream performance.

Figure 1 shows that the signals specialize. Answer confidence has AUROC 0.894 for pseudo-label correctness but 0.624 for the teacher-process target; teacher reliability has AUROC 0.649 and 0.788, respectively. A 10,000-replicate task-cluster bootstrap supports the separation without treating repeated rollouts as independent: for pseudo-label correctness, $C_A - R_T = +0.245$ (95% CI $[+0.076, +0.465]$, 57 tasks), while at the frozen 50% process-pass definition, $R_T - C_A = +0.164$ ($[+0.015, +0.383]$, 16 process-auditable tasks). At stricter 75% and 100% process-pass thresholds the $R_T > C_A$ point direction remains, but the paired difference intervals cross zero. Structural signals show the same descriptive dissociation: derivation-family count is 0.337 versus 0.709, and axiom-resolution is 0.607 versus 0.777. Appendix A reports the full uncertainty and threshold-sensitivity table.

3.3 Same-draft randomized correction tests are ceiling-limited, not evidence of equivalence

Prevalence and process-audit failure do not imply causal harm. The artifact contains two complete Gemini same-draft randomized interventions under the same design (40 drafts, 160 trials, seed

7). In the later run, process-correct versus Lucky-Correct teachers yield ACR 0.950 versus 0.9375 ($\Delta = +1.25\,pp$) and PCR 0.7375 versus 0.7625 ($\Delta = -2.5\,pp$). An earlier complete Gemini run yields ACR 0.9375 versus 0.900 ($\Delta = +3.75\,pp$) and PCR 0.750 versus 0.7625 ($\Delta = -1.25\,pp$). In both runs, answer correction is already 0.90–0.95 in each condition. This near-ceiling regime leaves little headroom for detecting answer-level harm, so the result is *inconclusive*: it does not establish equivalence or non-harm. A completed local Qwen3-4B sensitivity run has the opposite ACR direction (0.833 versus 0.917) and a small positive PCR difference (0.667 versus 0.625), reinforcing model/setting dependence rather than a universal effect. Appendix B discloses the recorded run ledger rather than selecting one run post hoc.

3.4 Selector rankings expose objective mismatch

A separate correction-time study contains 558 repeated trials over 12 unique tasks. No pairwise selector comparison against the longest-teacher reference survives multiplicity correction. The descriptive ranking itself is informative: random-teacher selection has nominal ACR 0.828 versus 0.804 for longest-teacher selection, while medoid has the highest PCR (0.804). The "oracle" arm uses the hidden *answer* only and is therefore renamed *answer-gold oracle*: it has high ACR (0.870) but PCR 0.609, below medoid. This is not a paradox; it shows that an answer oracle is not a process oracle. Unequal trial counts reflect condition availability/task coverage, so Table 9 is descriptive rather than a balanced-factorial causal ranking. Together these results motivate a control formulation in which the optimized target must be stated explicitly.

4 RAST: A Reliability-Control Layer

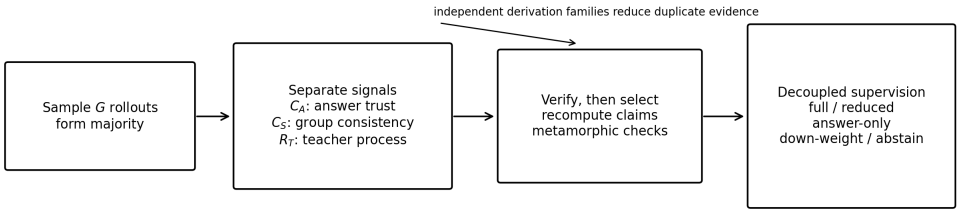

Figure 2: RAST separates answer, group, and teacher-level evidence before choosing how much of a self-generated trace to learn from. It is a control layer for self-distillation, not a claim that one scalar score certifies correctness.

For an unlabeled problem x, U-OPSD samples G rollouts, forms a majority-consistent pseudo-solution, conditions a teacher distribution on a short pseudo-solution, and distills it into prefixes of an incorrect on-policy completion (Li et al., 2026). RAST adds a decision layer with three distinct quantities. C_A describes answer support (vote share, margin, ties, validity, and optional effective-vote correction); C_S measures group trace consistency over aligned decision quantities; and R_T scores a candidate teacher using trace-level evidence such as self-consistency, independent derivation families, and problem-internal verification. Mechanical checks can recompute Bayesian quantities, execute SQL/code, or test plan constraints; metamorphic checks test invariances when exact recomputation is unavailable.

The action layer decouples answer weight w_A and trace weight w_T. It can apply full reasoning, reduced trace, answer-only supervision, down-weighting, or abstention. The full conceptual design also includes effective-vote and derivation-family components. Our experiments directly test signal separation, threshold risk–coverage, teacher-selector variants, verifier-passed selection, and weight-update approximations; they do not claim a causal ablation of every conceptual component. Appendix A gives the operational pseudocode and threshold sweep used in the diagnostic analysis.

5 EXPERIMENTAL PROGRAM

5.1 FOUR EVIDENCE LAYERS

We use four complementary layers rather than treating one training table as proof of the framework. **Diagnostic audit:** a 133-problem manufacturing reasoning suite supports answer/process audits; 29 diagnosis tasks contain process gold, with Gemini 2.5 Pro audited broadly and a separately collected Qwen3-4B process-gold subset used as the student-family bridge. **Correction studies:** repeated trials compare teacher selectors and same-draft randomized process-correct versus Lucky-Correct teachers. **Frozen weight study:** a separate protocol uses 125 training problems and a disjoint 125-problem hidden set, balanced across code, numeric, SQL, planning, and diagnosis (25 hidden problems each). **Public-benchmark downstream stress test:** after freezing those results, MATH-500 provides a downstream stress test on 500 public mathematical-reasoning tasks. Hidden labels and process gold are audit/evaluation-only and never enter pseudo-label construction or training.

5.2 WEIGHT-STUDY MODEL AND STACKS

The student is a Qwen3-4B-family model adapted with LoRA on the final eight layers (rank 8, scale 20). Sampling uses $G = 8$, temperature 0.7, max 2048 generated tokens, and per-rollout seeds. Seeds 1–3 use `mlx-community/Qwen3-4B-4bit` with MLX/Metal; seeds 4–5 use a Hugging Face bf16/CUDA stack. The algorithm version, prompt, and dataset are frozen (uopsd-v1.4; git `d244e413`; prompt hash `9c03b8c3c1c91dc2`; dataset hash `9cd761bb0006e512`). Because backend/precision and seed group are confounded, we report pooled five-seed sensitivity results and the homogeneous three-seed MLX subset separately. A post-freeze between-seed sensitivity analysis for the strongest point-estimate variant is reported in Appendix D; it does not support a winner claim.

5.3 COMPARED VARIANTS AND VALIDITY

Each seed contains baseline/no training, a *legacy ungated U-OPSD-style reference* (the frozen local arm uses no reliability gate, a longest-agreeing teacher heuristic, and up to four minority trajectories), gated U-OPSD, answer-only, vote weighting, medoid teacher selection, strict reliability gating, and answer–teacher gating. The legacy reference is retained because it was part of the frozen 40-job design, but we no longer describe it as an exact reproduction of the source-paper recipe. All final training jobs changed adapter weights relative to the common initialization. Before final analysis, 56 historical runs were invalidated and retained in audit history (degenerate sampling, zero-adapter no-op, interrupted jobs, mock contamination, or baseline inconsistency). Final status is 40/40 done, 40 valid, zero invalid pending.

5.4 STATISTICAL PLAN

The weight-study primary metric is hidden-set Majority@8 change, adjusted by the same seed's baseline re-measurement change. Pooled bootstrap intervals resample the task–seed structure. The pre-specified multiplicity family contains six proposed intervention variants—gated U-OPSD, answer-only, vote weighting, medoid, reliability gating, and answer–teacher gating—each evaluated against the no-training baseline. The legacy ungated reference is reported outside that Holm family. Because that arm was already frozen and executed before the source-paper fidelity audit, changing it would mix protocols; we therefore preserve it as historical confirmatory context rather than reinterpret it post hoc. Baseline re-measurement noise has mean absolute Majority@8 change 1.44 points (SD 2.06).

5.5 POST-FREEZE U-OPSD BASELINE-FIDELITY CHECK

After freezing the primary 40-job analysis, we ran a separate HF/CUDA two-seed fidelity check that more closely follows the audited U-OPSD structure: $G = 8$, $\tau = 0.5$, shortest agreeing pseudo-solution, one seeded disagreement trajectory, fixed initial-policy teacher, forward KL, and no RAST gate. The shared Qwen3-4B LoRA/optimizer/evaluation stack is otherwise unchanged, so this is descriptive robustness evidence rather than an exact source-paper reproduction; Appendix E gives the full protocol and provenance audit.

5.6 POST-FREEZE MATH-500 PUBLIC-BENCHMARK DOWNSTREAM STRESS TEST

We additionally froze a public-benchmark downstream stress test on MATH-500, a 500-problem subset of the MATH benchmark (Hendrycks et al., 2021; Lightman et al., 2023). A deterministic 256-item MATH training subset, disjoint from the 500 evaluation prompts by normalized problem hash, supplies only prompts/public metadata to self-distillation; reference solutions and answers are withheld from collection and training. Seeds 6–8 use Qwen3-4B with the HF bf16/CUDA stack, exact M0 reload before each learned condition, $G = 8$, temperature 0.7, max 2048 tokens, forward-KL top-$k = 100$, and one-pass LoRA training. We compare a *structure-faithful implementation under our frozen training regime* against the frozen answer–teacher gate. It is not an exact reproduction of the source U-OPSD recipe: our implementation shares the study LoRA/optimizer stack, uses temperature 0.7 and a 2048-token cap, a top-$k = 100$ KL approximation, no pointwise KL clipping, and one-pass training. Math-Verify canonicalizes symbolic answers; tied majority votes count as abstention/incorrect. This study was specified after the primary family was frozen and is therefore secondary downstream stress-test evidence, not part of the original Holm family. Uncertainty uses 10,000 paired crossed bootstrap replicates that resample the three seeds and 500 task IDs independently while preserving within-cell condition pairing; Appendix F gives protocol, seedwise results, and provenance.

6 RESULTS

6.1 FIVE-SEED WEIGHT-LEVEL RESULTS

Table 3 reports the frozen weight study. Answer–teacher gating has the largest baseline-adjusted point estimate (+2.72 points), but its 95% CI crosses zero and the bootstrap p-value is 0.0725. After Holm correction, none of the six primary intervention comparisons is significant. The paired McNemar $p = 0.0293$ is exploratory and does not override the non-significant primary multiplicity-corrected analysis.

Relative to the legacy ungated reference, answer–teacher gating is +1.44 points on average across the five paired seed deltas. A post-hoc seed-paired t-test gives $p = 0.266$ (95% CI $[-1.65, +4.53]$) and Wilcoxon $p = 0.375$. This direct comparison therefore also fails to support a superiority claim.

6.2 POST-FREEZE BASELINE-FIDELITY RESULT

The structurally closer baseline has raw Majority@8 changes of +0.8 and +2.4 points in seeds 4 and 5. Relative to the already-frozen same-seed baseline re-measurements (-0.8 and +3.2), the corresponding descriptive net changes are +1.6 and -0.8 points, for a mean of +0.4. Thus the raw changes are non-negative in both replications, but the two-seed effect is of the same order as the 1.44-point mean absolute re-measurement change and does not support equivalence or superiority. No inferential test is attached; Appendix E reports the secondary metrics and exact audit trail.

6.3 MATH-500 DOWNSTREAM STRESS-TEST RESULT

Across the three frozen MATH-500 seeds, M0 averages 81.40% Single, 86.80% Majority@8, and 93.33% Pass@8. The paper-structure-faithful U-OPSD condition averages 81.20/85.80/92.87%, while the RAST answer–teacher gate averages 81.93/85.73/92.80%. Relative to same-seed M0, the Majority@8 changes are -1.00 pp for U-OPSD (95% CI $[-2.47, +0.60]$) and -1.07 pp for RAST ($[-2.53, +0.40]$); the direct RAST–U-OPSD contrast is -0.07 pp ($[-1.40, +1.33]$). RAST has a small descriptive Single shift of +0.53 pp versus M0 ($[-1.53, +2.60]$), but this interval also crosses zero. Invalid-output rate is 0 in all conditions. Tie rate rises descriptively from 3.73% at M0 to 5.53% for U-OPSD and 5.20% for RAST; because ties were pre-specified as abstention/incorrect, this accompanies the lower Majority@8 values without establishing causation. The replication therefore does not identify a downstream accuracy winner and narrows the downstream-transfer claim rather than reversing the reliability findings. Full seedwise values and bootstrap contrasts are in Appendix F.

Table 3: Pooled five-seed Majority@8 results. Net gain subtracts within-seed baseline re-measurement. CI and p_{boot} use task–seed cluster bootstrap. Values are percentage points.

Variant	Mean Δ	Net Δ	95% CI	p_{boot}	Holm p
Answer–teacher gate	2.56	**2.72**	$[-0.16, 5.76]$	0.0725	0.435
Vote weight	2.40	2.56	$[-0.64, 5.76]$	0.1245	0.6225
Medoid	2.08	2.24	$[-0.64, 5.12]$	0.1540	0.6225
U-OPSD (gated)	2.08	2.24	$[-0.80, 5.28]$	0.1730	0.6225
Answer only	1.60	1.76	$[-1.60, 5.12]$	0.3630	0.7260
Legacy ungated ref.	1.12	1.28	$[-1.44, 4.00]$	0.3985	–
Reliability gated	0.64	0.80	$[-2.24, 3.68]$	0.6655	0.7260

6.4 Coverage, volume, and instability

The weight results match the diagnostic risk–coverage story qualitatively. The legacy ungated reference accepts 51–66 distillation examples in the MLX seeds, more than twice the 22–31 examples used by several gated variants, yet its pooled net gain is only +1.28 points. Strict reliability gating retains only 11–15 examples and is weakest (+0.80). More supervision and stricter filtering can both underperform. Vote weighting has the second-largest point estimate (+2.56) but the clearest secondary-metric failure: in seed 2, Pass@8 falls 0.832→0.712 while invalid-output rate rises 0.026→0.063. The final 40-job export now permits complete five-seed secondary analysis. Baseline-adjusted mean net changes show that answer–teacher gating improves Single by +2.56 points and Pass@8 by +1.60 while slightly reducing invalid output (-0.20); medoid improves Single/Pass@8 by +1.60/+1.92 and reduces invalid output by -0.50. By contrast, vote weighting improves Single by +1.92 but reduces Pass@8 by -1.12 and increases invalid output by +0.38. These results reinforce the multi-objective interpretation: a variant can improve majority or single-sample accuracy while worsening coverage or validity. Appendix D reports the complete table.

A provenance audit shows that vote weighting and answer–teacher gating collapse to identical optimizer-facing sample weights in seeds 1 and 3 despite distinct run IDs and different raw signal values; those seeds therefore provide no contrast between the variants. Appendix D.2 records the aligned-example counts, hashes, and the limit of what the archived gate state can identify.

6.5 Homogeneous subset and domain heterogeneity

Within the three-seed MLX subset, answer–teacher gating has a larger baseline-adjusted gain (+4.27 points; 95% CI $[+0.53, +8.27]$; $p_{boot} = 0.034$), but Holm-adjusted $p = 0.204$. The pooled estimate is smaller, illustrating how a three-seed slice can overstate evidence. In the MLX subset, gains concentrate in SQL and variably in planning/diagnosis; code remains fixed at 0.80 Majority@8 across local jobs and numeric changes are comparable to baseline noise. The code ceiling may reflect task/evaluator saturation rather than absence of any possible training effect; we therefore avoid a domain-general claim.

7 Discussion

What is supported. The strongest positive evidence is not a winner algorithm but a separation of targets. Lucky Correct occurs in both the Gemini audit arm and an explicitly identified Qwen3-4B process-gold audit subset. Answer confidence and teacher reliability predict different audit targets. A joint gate traces a risk–coverage frontier below the expected audit-defined risk of random thinning from the same candidate pool. Finally, correction-time selector rankings differ depending on whether answer correction or process correction is optimized; even an answer-gold oracle is not a process oracle.

What remains unresolved. The correction interventions do not establish that process-wrong teachers cause downstream harm. Their Gemini ACR is near ceiling, and the local sensitivity run changes the sign of the answer-level contrast. Likewise, no selector wins after multiplicity correction and no weight-update variant is superior after Holm correction. The MATH-500 downstream

evaluation also leaves downstream accuracy unresolved: both self-distillation conditions are about one point below M0 on Majority@8, and their direct difference is near zero with an interval spanning effects in either direction. These findings should not be read as equivalence tests; the experiments remain underpowered for small training effects and, in the correction setting, structurally insensitive to some forms of harm.

Why a control layer is still useful. Answer and process objectives can rank selectors differently, while strict gating can reduce risk by discarding supervision. The MATH-500 result makes the scope explicit: separating trust signals is a measurement/control claim, not a guarantee that every retained supervision policy improves benchmark accuracy. RAST therefore treats self-teaching as a multi-objective coverage/risk control problem, not a scalar correctness certificate.

Learned PRMs. When step labels or a trusted judge exist, learned process reward models are stronger alternatives (Lightman et al., 2023; Wang et al., 2023b); we study the narrower no-judge regime where problem-internal checks provide an auditable but domain-dependent fallback.

Statistical power and detectability. With 125 hidden problems, one item is 0.8 points; the +2.72-point primary estimate is only about 3–4 net items and is comparable to baseline re-measurement noise. Five training seeds are few. A deliberately conservative seed-as-unit sensitivity analysis gives $t(4) = 1.39$, $p = 0.236$, 95% interval $[-2.70, +8.14]$ pp, Wilcoxon $p = 0.25$, and a 100,000-replicate seed-bootstrap interval $[-0.96, +5.92]$ pp (Appendix D). Thus the study is better at rejecting large winner claims than resolving small effects.

8 Limitations

Process auditing is selective: Gemini has 28.5% measurable rollouts, whereas the Qwen bridge deliberately samples process-gold tasks and therefore has 232/232 measurable traces by construction. Conditional Lucky-Correct rates are not population prevalence estimates and should not be compared across models. Common-mode error lacks a frozen prevalence label; the selector study has only 12 unique tasks; and the same-draft Gemini interventions are near ceiling, so they do not support equivalence or universal non-harm. The pooled weight study confounds seed group with execution stack (MLX 4-bit/Metal versus HF bf16/CUDA), uses only five training seeds, one Qwen3-4B-family student, and a constructed manufacturing-reasoning suite. We now include an author-run MATH-500 downstream evaluation, but external validity is still limited to one public mathematical-reasoning benchmark, one model family, three public-benchmark seeds, and a shared one-pass LoRA/optimizer stack. That evaluation does not resolve a Majority@8 improvement for either self-distillation condition relative to M0, so broader replication across models, domains, and training budgets is required before making downstream-accuracy claims. Mechanical verification is domain-dependent, and some conceptual components are not isolated causally; in particular, C_S is a group-level diagnostic/control signal rather than a separately randomized factor in the final weight study. The anonymous package exactly specifies R_T and the action/weight map, but bit-for-bit regeneration of the archived raw-rollout C_A/C_S constructors still requires the full repository, so we do not claim end-to-end executable reproduction from the submission snapshot alone. Multiplicity control applies only to the original primary Majority@8 family; the post-freeze fidelity and MATH-500 studies are secondary/descriptive and are not folded into that Holm family.

9 Conclusion

Consensus is useful, but it certifies less than self-distillation often asks of it. Answer-correct traces can fail process audit, answer confidence and teacher reliability predict different targets, and selector rankings depend on the protected objective. Neither the frozen weight study nor the post-freeze MATH-500 downstream stress test establishes a downstream accuracy winner. Our claim is therefore not that U-OPSD fails or that reliability gating guarantees higher benchmark accuracy, but that aggregate answer behavior does not certify the trace being learned. Self-distillation should monitor answer trust, process trust, and supervision coverage separately rather than collapse them into one consensus score.

AI USE STATEMENT

Generative AI tools, including ChatGPT, were used for research ideation and framing, methodology feedback, implementation/debugging assistance, interpretation of experimental outputs, literature discovery, drafting/editing, and preparation of tables/figures. All experiments and validation scripts were executed in author-controlled environments; reported numerical results were taken from frozen artifacts and validation outputs rather than generated by an LLM. AI-assisted code, statistical claims, citations, figures, and manuscript text were manually reviewed. No generative AI tool was used to fabricate measurements or modify hidden evaluation labels. The authors take responsibility for the final submission.

ETHICS STATEMENT

The experiments use constructed reasoning problems and do not involve human subjects, personal data, or sensitive user records. Hidden answers are evaluation/audit-only. Because self-improvement can amplify errors, the paper emphasizes abstention, multi-metric monitoring, and explicit separation of answer trust from trace trust rather than presenting autonomous self-training as intrinsically safe.

REPRODUCIBILITY STATEMENT

The frozen weight-study version, prompt hash, dataset hash, common initialization fingerprint, per-seed configuration, validation status, statistical protocol, and invalidated-run history are recorded. The supplementary material reports operational pseudocode, literal v1.4 control constants, an executable reference implementation of the documented action/weight map, threshold sweeps, task-cluster diagnostic uncertainty/sensitivity tables, selector/correction tables, full seed-wise weight results, complete five-seed secondary metrics, and audit history. The post-freeze baseline-fidelity protocol is preserved separately. For MATH-500, the supplement records the frozen 256-item collection pool and 500-item evaluation hashes, seeds 6–8, exact M0 fingerprint, condition definitions, per-seed metrics/adapters, the analysis-plan lock, and 10,000-replicate paired two-way bootstrap summary; the original 40-job result remains byte-identical. The operational supplement gives the selected-teacher scorer, action/weight mapping, admission constants, and the exact reduced-trace scale (0.30). It also contains an executable anonymous core-reference archive for R_T and the control/weight map, a component-by-component reproducibility matrix, and a source-evidence map that records the exact partial provenance available for C_A/C_S without reverse-engineering missing formulas. A descriptive MATH tie-excluded sensitivity is supplied separately and does not redefine the frozen all-task metric. This core archive is not the full experiment repository. A full anonymous executable-code export remains an author-side release gate: without it, we explicitly do not claim bit-for-bit independent recomputation of the archived C_A/C_S scalar constructors from this supplementary package alone.

REFERENCES

Y. Li, B. Wang, Y. Liang, Y. Tian, D. Fu, and N. Vasconcelos. On-Policy Self-Distillation without Any Supervision. arXiv:2608.06296, 2026.

S. Zhao, Z. Xie, M. Liu, J. Huang, G. Pang, F. Chen, and A. Grover. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models. arXiv:2601.18734, 2026.

X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. ICLR, 2023.

M. Prabhudesai, L. Chen, A. Ippoliti, K. Fragkiadaki, H. Liu, and D. Pathak. Maximizing Confidence Alone Improves Reasoning. arXiv:2505.22660, 2025.

A. Prasad, W. Yuan, R. Y. Pang, et al. Self-Consistency Preference Optimization. OpenReview, 2025.

W. Yuan, R. Y. Pang, K. Cho, X. Li, S. Sukhbaatar, J. Xu, and J. Weston. Self-Rewarding Language Models. arXiv:2401.10020, 2024.

E. Zelikman, Y. Wu, J. Mu, and N. D. Goodman. STaR: Bootstrapping Reasoning with Reasoning. NeurIPS, 2022.

J. Huang, S. Gu, L. Hou, Y. Wu, X. Wang, H. Yu, and J. Han. Large Language Models Can Self-Improve. EMNLP, 2023.

S. Shafayat, F. Tajwar, R. Salakhutdinov, J. Schneider, and A. Zanette. Can Large Reasoning Models Self-Train? arXiv:2505.21444, 2025.

R. Agarwal, N. Vieillard, Y. Zhou, P. Stanczyk, S. Ramos, M. Geist, and O. Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. ICLR, 2024.

Y. Gu, L. Dong, F. Wei, and M. Huang. MiniLLM: Knowledge Distillation of Large Language Models. ICLR, 2024.

J. Ko, S. Kim, T. Chen, and S.-Y. Yun. DistiLLM: Towards Streamlined Distillation for Large Language Models. ICML, 2024.

Y. Fu, H. Huang, K. Jiang, Y. Zhu, and D. Zhao. Revisiting On-Policy Distillation: Empirical Failure Modes and Simple Fixes. arXiv:2603.25562, 2026.

R. Wang, H. Wang, Y. Chen, B. Xue, T. Fang, W. Yu, and K.-F. Wong. Demystifying On-Policy Distillation: Roles, Pathologies, and Regulations. arXiv:2607.13399, 2026.

J. Ke, Z. Wen, W. Li, C. He, and L. Zhang. Respecting Self-Uncertainty in On-Policy Self-Distillation for Efficient LLM Reasoning. arXiv:2605.13255, 2026.

X. Liu, X. Wang, Y. Ma, Y. Zhang, and C. Xiao. When Are Teacher Tokens Reliable? Position-Weighted On-Policy Self-Distillation for Reasoning. arXiv:2605.21606, 2026.

Z. Tan and Y. Hong. Self-Supervised On-Policy Distillation for Reasoning Language Models. arXiv:2605.17497, 2026.

H. Lightman, V. Kosaraju, Y. Burda, et al. Let's Verify Step by Step. arXiv:2305.20050, 2023.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. NeurIPS, 2021.

P. Wang, L. Li, Z. Shao, et al. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. arXiv:2312.08935, 2023.

M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. NeurIPS, 2023.

T. Lanham, A. Chen, A. Radhakrishnan, et al. Measuring Faithfulness in Chain-of-Thought Reasoning. arXiv:2307.13702, 2023.

T. Y. Chen, S. C. Cheung, and S. M. Yiu. Metamorphic Testing: A New Approach for Generating Next Test Cases. Technical Report HKUST-CS98-01, 1998.

S. Segura, G. Fraser, A. B. Sanchez, and A. Ruiz-Cortes. A Survey on Metamorphic Testing. IEEE Transactions on Software Engineering, 42(9), 2016.

A. Stein, F. Huang, and T. Goldstein. GATES: Self-Distillation under Privileged Context with Consensus Gating. arXiv:2602.20574, 2026.

Y. Geifman and R. El-Yaniv. Selective Classification for Deep Neural Networks. arXiv:1705.08500, 2017.

A Operationalization and Threshold Sweeps

A.1 RAST control logic

The implementation separates three signals and two weights. The following pseudocode states the decision structure without conflating the diagnostic threshold sweep with the frozen weight-update variants:

```
rollouts = sample(model, x, G)
a = majority(canonicalize(answer(rollouts)))
candidates = rollouts supporting a
CA = answer_confidence(rollouts, a)
CS = group_trace_consistency(candidates)
for candidate in candidates:
    RT[candidate] = teacher_reliability(candidate,
        independent_derivations, mechanical_checks,
        metamorphic_checks)
selector = selector_for_variant(variant)
teacher = selector(candidates, RT)
policy = supervision_policy_for_variant(variant)
(wA, wT, action) = policy(CA, CS, RT[teacher])
update(answer_weight=wA, trace_weight=wT, action=action)
```

Longest, medoid, random, reliability-selected, and verifier-passed variants replace the selector function; answer-only, weighting, strict gating, and joint answer–teacher variants replace the supervision policy and/or admission thresholds. The diagnostic risk–coverage analysis sweeps thresholds $\{0.30, 0.45, 0.60, 0.75\}$ rather than choosing a post-hoc optimum. For the joint answer–teacher gate, (coverage, audit-defined risk) is (0.7341, 0.2287), (0.3527, 0.1021), (0.1961, 0.0487), and (0.0972, 0.0045). The full-coverage candidate pool has audit-defined risk 0.3053, so random thinning independent of the audit label has expected risk 0.3053 at every coverage; Figure 1 shows this expectation as a dashed reference rather than as a simulated control. Frozen machine-readable artifacts retain the historical label `harmful_update`; the manuscript uses "audit-risk event" to avoid implying a causal downstream effect. For reproducibility, we state the frozen teacher scorer and action/weight map directly. Let S_i be the set of decisive numeric quantities in agreeing trace i after removing numbers already present in the prompt. For a selected teacher t, let P contain quantities appearing in at least $\max(2, \lceil |A|/2 \rceil)$ agreeing traces, where A is the agreeing set. Define $F_t = F_1(S_t, P)$ when P is nonempty, and J_t as the mean Jaccard overlap between S_t and the other agreeing traces. The representativeness term is the mean of the available F_t and J_t values (0.5 if neither is measurable). A deterministic trace checker gives self-consistency $q_t \in [0, 1]$; an uncheckable trace is assigned 0.3, while an unsafe/internally contradictory trace multiplies this value by 0.4. The depth gate is $d_t = \min(1, |S_t|/3)$; detected degeneracy multiplies d_t by 0.3, as does violation of an applicable problem-internal normalization axiom. For traces with decisive quantities,

$$R_T(t) = d_t \, [0.55 \, \mathrm{repr}_t + 0.45 \, q_t].$$

For domains with no extracted decisive quantities, the available public mechanical-check score is used directly (otherwise 0). The reliability selector chooses the agreeing trace with maximal R_T, ties broken by rollout order. Hidden answers and process gold are not inputs to this score.

The frozen v1.4 action layer uses C_A for answer trust and selected R_T for trace trust. It first rejects ties/non-admissible rows and rows with no minority trajectory. The archived source-level defaults are now stated literally in Table 4: eligibility and valid-rate thresholds 0.50, vote-margin threshold 0.25, answer/support thresholds 0.25, trace-support threshold 0.25, reliability floor 0.45, and reduced-trace scale 0.30. The action breakpoints are $C_A = 0.25$ (hard abstention), $C_A = 0.60$ (answer-support breakpoint), and $R_T \in \{0.65, 0.80\}$ (reduced/full trace bands; legacy trace-confidence fallback bands are 0.35/0.45). An unresolved scale-family split suppresses full trace; an unsupported trace becomes answer-only. The two weights obey $w_T \leq w_A$: abstain/need-more/no-update sets both to zero; answer-only uses $(w_A, w_T) = (C_A, 0)$; downweight uses $w_A = 0.5C_A$ and $w_T = 0.5 \min(w_A, R_T)$; full reasoning uses $(C_A, \min(C_A, R_T))$; reduced trace uses $w_A = C_A$ and $w_T = 0.30 \min(C_A, R_T)$. The supplement includes a small executable reference implementation of this signal-to-action/weight map; it is not represented as the full training repository.

Frozen source key	Literal value	Role in v1.4 snapshot
`eligibility_tau`	0.50	answer-consensus admission floor
`valid_rate_tau`	0.50	valid-rollout-rate floor
`margin_tau`	0.25	vote-margin guard
`answer_teacher_tau`	0.25	joint answer/teacher admission constant
`answer_support_tau`	0.25	answer-support guard
`trace_support_tau`	0.25	trace-support guard
`reliability_floor`	0.45	reliability-gated variant floor
`reduced_trace_scale`	0.30	multiplier for reduced trace weight
`tie_policy`	abstain	no teacher on tied top answer
`empty_minority_policy`	no-uopsd	no update when $Y^- = \varnothing$
`half_vote_policy`	adaptive-G	ambiguous half-vote handling

Table 4: Literal v1.4 defaults recovered from the frozen source/version manifest. These constants are reported as implementation provenance; not every key is an independent tuned parameter in every variant.

C_A and C_S remain separate frozen implementation signals rather than being re-fit from hidden outcomes. C_A is the archived answer-support aggregate over vote support/margin/validity (with optional effective-vote correction); C_S summarizes agreement of decisive quantities among answer-supporting traces. The preserved source evidence additionally shows that the group-support/minority quantity sets remove numbers already present in the prompt, invokes an independence correction after the group signal is formed, and records an alternate effective-vote scalar as $C_A^{\text{eff}} = (f + m)/2$, where f and m are that routine's effective-vote fraction and margin. However, the inner component list used by the archived constructor of the primary C_A scalar is not present in the submission-source snapshot. We therefore do not reverse-engineer a replacement closed form from hidden labels. The diagnostic conclusions consume the frozen scalars with task-cluster resampling. To make this boundary auditable rather than implicit, Table 5 separates the components that are exactly independently implementable from those that still require the full repository; the supplementary file `RAST_SIGNAL_PROVENANCE_v8.md` enumerates the recovered source-line evidence. The anonymous supplement also includes executable reference code for the exact R_T scalar algebra and for the action/weight policy. These files are deliberately labeled as core reference code, not as a substitute for the full experiment repository.

Component	Status	Independent reconstruction from submitted package
Selected-teacher R_T algebra	exact	Yes after decisive-quantity/public-check extraction; executable reference supplied.
Action ordering / thresholds	exact recovered	Yes for already-computed signals; literal constants and reference policy supplied.
Two-level w_A, w_T map	exact	Yes; including reduced-trace factor 0.30 and $w_T \le w_A$.
Archived C_A, C_S analyses	frozen scalars	Reported analyses can be re-audited from released values/summaries.
Raw-rollout C_A, C_S constructors	incomplete snapshot	No bit-for-bit regeneration from this package alone; no formula is inferred from hidden labels.
Full end-to-end repository	not included	Remains the release gate for complete environment/helper reconstruction.

Table 5: Reproducibility completeness of the anonymous submission package. "Exact" refers to components whose frozen algebra/branching is supplied; it does not imply that missing raw-rollout constructors have been recreated.

Target	Process pass	C_A AUROC [95% CI]	R_T AUROC [95% CI]	Paired AUROC difference [95% CI]
Answer correctness	–	0.894 [0.807,0.978]	0.649 [0.457,0.794]	$C_A - R_T$: +0.245 [0.076,0.465]
Process audit	50%	0.624 [0.368,0.802]	0.788 [0.618,0.899]	$R_T - C_A$: +0.164 [0.015,0.383]
Process audit	75%	0.727 [0.402,0.854]	0.850 [0.667,0.922]	$R_T - C_A$: +0.123 [-0.017,0.361]
Process audit	100%	0.743 [0.435,0.939]	0.878 [0.671,0.959]	$R_T - C_A$: +0.135 [-0.111,0.366]

Table 6: Task-cluster uncertainty and process-pass sensitivity. Intervals use 10,000 bootstrap replicates over task IDs. The answer analysis has 310 repeated rows from 57 tasks; process analyses have 110 rows from 16 process-auditable tasks. The 75% and 100% rows are sensitivity analyses, not alternative frozen selection rules.

Threshold	Frozen joint (coverage, risk)	Reconstructed coverage [95% CI]	Reconstructed risk [95% CI]
0.30	(0.7341, 0.2287)	0.7344 [0.6538,0.8101]	0.2317 [0.1418,0.3360]
0.45	(0.3527, 0.1021)	0.3515 [0.2565,0.4503]	0.0981 [0.0193,0.2161]
0.60	(0.1961, 0.0487)	0.1948 [0.1237,0.2719]	0.0391 [0.0084,0.0945]
0.75	(0.0972, 0.0045)	0.0985 [0.0509,0.1551]	0.0045 [0.0000,0.0183]

Table 7: Risk–coverage uncertainty sensitivity. The first column pair is the frozen summary used for the manuscript point estimates. The bootstrap columns independently reconstruct the joint gate from archived row-level answer-confidence/teacher-reliability fields and resample 100 task clusters 10,000 times. Reconstruction differs from the frozen points by at most 0.13 percentage points in coverage and 0.96 points in risk, so the intervals are used as an uncertainty sensitivity rather than to redefine the frozen estimand.

A.2 Variant definitions and observed coverage

Variant	Operational change	MLX retained examples
Baseline	No training; repeat evaluation for noise estimate	0
Legacy ungated ref.	Frozen local code: no gate; longest teacher; up to 4 minority trajectories	51–66
Gated U-OPSD	Common admission gate before distillation	22–31
Answer only	Suppress full-trace imitation	22–31
Vote weight	Scale supervision by consensus strength	22–31
Medoid	Select most central agreeing trace	22–31
Reliability gated	Strict trace admission	11–15
Answer–teacher gate	Joint answer/trace admission	22–31

Table 8: Variant operationalization. Reported retained counts are observed final-run quantities, not hand-tuned targets; Appendix A and the machine-readable operational specification document the frozen control logic and known scalar breakpoints.

B Selector and Randomized-Teacher Studies

B.1 Direct selector study

The correction study contains 558 repeated trials over 12 unique tasks. Arm counts differ because not every condition is available on every recorded task/experiment; therefore the aggregate rates are descriptive; pairwise ACR/PCR contrasts use task-cluster bootstrap with Holm correction rather than treating trials as independent. Table 9 reports the rates.

Condition	n	ACR	PCR
No teacher / resample	56	0.661	0.643
Self-reflection	68	0.500	0.518
Answer only	56	0.518	0.571
Longest-teacher ref.	56	0.804	0.643
Medoid	68	0.853	0.804
Reliability selected	56	0.696	0.714
Verifier passed	56	0.786	0.679
Random teacher	64	0.828	0.615
Process selected	32	0.750	0.625
Answer-gold oracle	46	0.870	0.609

Table 9: Direct correction-time teacher-selection study. Random teacher nominally exceeds longest-teacher ACR, while medoid has higher PCR; the answer-gold oracle is an answer ceiling, not a process oracle. No pairwise comparison against the longest-teacher reference survives multiplicity correction.

B.2 Recorded same-draft randomized teacher interventions

The artifact contains every run shown in Table 10; the main text does not select one run because it is "latest." The two complete Gemini runs are comparable (same design, seed, and 640-token setting) and both operate near an answer-correction ceiling. The completed local run uses a different model and 2048-token limit and is treated as sensitivity evidence. An earlier completed local record has all-zero outcomes in both conditions; it is excluded from pooled inference because provenance is insufficient at the analysis freeze to distinguish pipeline failure from a genuine degenerate run. One additional local run is also excluded because it was incomplete at the freeze.

Run	Arm / state	Trials	ΔACR	ΔPCR	Note
E3-20b27046	Gemini / complete	160	+1.25pp	-2.50pp	ACR 0.95/0.938
E3-6e2f224c	Gemini / complete	160	+3.75pp	-1.25pp	ACR 0.938/0.900
E3-42f64c69	Qwen / complete	96	-8.34pp	+4.17pp	sensitivity; 2048 tokens
E3-2a9a5da2	Qwen / excluded	84	0.00pp	0.00pp	all-zero completed record; provenance unresolved
E3-99a3b455	Qwen / excluded	12	0.00pp	-16.66pp	incomplete at freeze; not pooled

Table 10: Run ledger for recorded same-draft randomized teacher experiments. Δ is process-correct minus Lucky-Correct. The first two rows are the complete Gemini runs summarized in the main text.

C Frozen Weight-Study Configuration

Item	Frozen setting
Student	Qwen3-4B family, LoRA
Local seeds 1–3	`mlx-community/Qwen3-4B-4bit`; MLX 0.32.0 / mlx_lm 0.31.3; Metal
Cloud seeds 4–5	Hugging Face stack; bf16; CUDA
LoRA	final 8 layers; rank 8; scale 20; 112 tensors
Sampling	$G = 8$; temperature 0.7; max_tokens 2048; per-rollout seeds
Train / hidden	125 training problems / 125 disjoint hidden problems
Frozen identifiers	uopsd-v1.4; git `d244e413`; prompt hash `9c03b8c3c1c91dc2`; dataset hash `9cd761bb0006e512`
Final validity	40/40 done; 40/40 valid; 0 invalid pending

The frozen 40-job *legacy ungated reference* used no reliability gate, the local longest-agreeing-teacher heuristic, and up to four minority trajectories per eligible task. We retain its results unchanged for the confirmatory audit trail, but do not label that arm an exact source-paper reproduction. The later structural-fidelity check is reported separately in Appendix E.

D Full Seed-Wise Weight Results

D.1 Between-seed sensitivity for the strongest point estimate

The answer–teacher gate has baseline-adjusted Majority@8 changes of $[+8.0, +4.8, 0.0, +4.0, -3.2]$ percentage points across seeds 1–5. Treating the five seeds as the resampling unit is deliberately conservative about between-seed variability: the mean is +2.72 pp, the sample SD is 4.37 pp, a two-sided one-sample t test gives $t(4) = 1.39$, $p = 0.236$, with 95% t interval $[-2.70, +8.14]$ pp; a two-sided Wilcoxon signed-rank sensitivity gives $p = 0.25$; and a 100,000-replicate seed bootstrap (RNG seed 20260923) gives percentile 95% interval $[-0.96, +5.92]$ pp. These post-freeze checks are not added to the original Holm family and do not change the conclusion that no accuracy winner is established. With only five seeds, and backend confounded with seed group, the interval is expected to be broad.

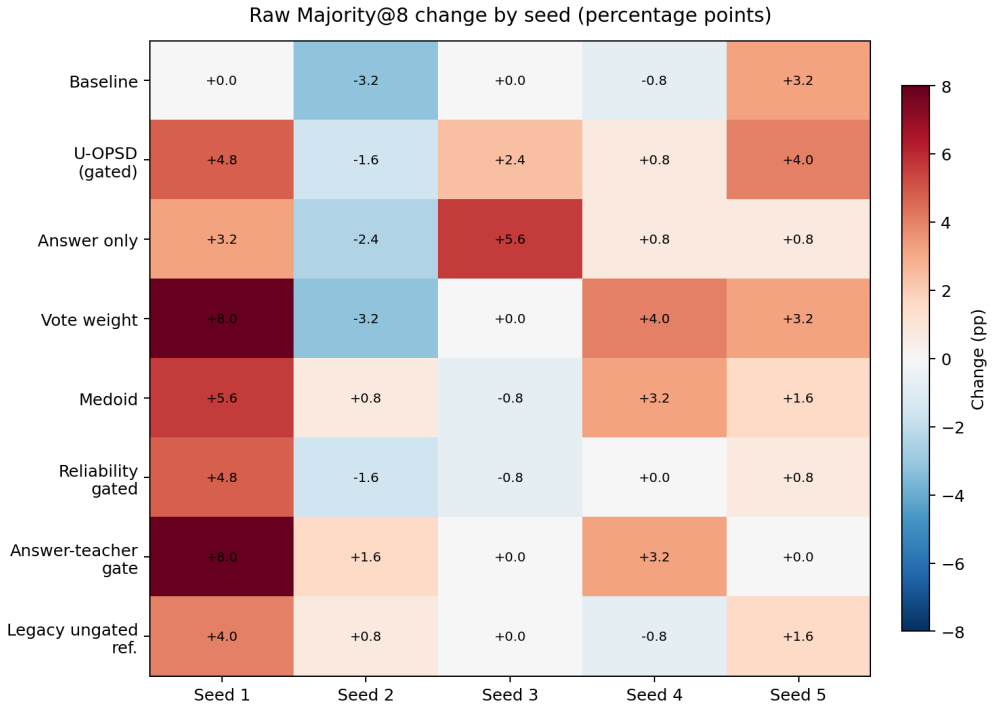

Figure 3: Raw Majority@8 change by method and seed. Values are percentage points. Vote weight and answer–teacher gate collapse to identical effective updates in seeds 1 and 3; Appendix D.2 gives the provenance audit.

Variant	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5
Baseline	0.0	-3.2	0.0	-0.8	+3.2
U-OPSD gated	+4.8	-1.6	+2.4	+0.8	+4.0
Answer only	+3.2	-2.4	+5.6	+0.8	+0.8
Vote weight	+8.0	-3.2	0.0	+4.0	+3.2
Medoid	+5.6	+0.8	-0.8	+3.2	+1.6
Reliability gated	+4.8	-1.6	-0.8	0.0	+0.8
Answer–teacher gate	+8.0	+1.6	0.0	+3.2	0.0
Legacy ungated ref.	+4.0	+0.8	0.0	-0.8	+1.6

Table 11: Raw Majority@8 change (percentage points) for every final seed.

D.2 Variant-collapse provenance audit

The exact equality of vote weighting and answer–teacher gating in seeds 1 and 3 is not a copied result. Their train/round identifiers differ, while the audit export aligns all 31 examples in seed 1 and all 29 in seed 3. The raw signal value differs in 19/31 and 15/29 aligned examples, respectively, but the gate-produced answer weight and logged applied sample weight match exactly for every aligned example (maximum absolute difference 0). Thus the collapse is not caused by rounding equal diagnostic signals. It is localized to the signal-to-sample-weight policy: the optimizer-facing sample weight equals the gate-produced `answer_weight`, and this quantity is exactly identical across the two variants for every aligned example in seeds 1 and 3. Because the archived audit does not retain every intermediate branch of the gate, we do not ascribe the equality to a specific clip or threshold operation. Once the applied weights coincide, token counts, loss, and final adapter hash are identical.

Seed	Examples	Tokens	Mean loss	Final adapter SHA-256 (prefix)
1 (both variants)	31	18,617	0.077085	`eae6f11a5959...`
3 (both variants)	29	14,717	0.164658	`a929b7e843a8...`

Table 12: Provenance check for the two seeds in which vote weighting and answer–teacher gating collapse to the same effective update. Distinct train IDs and round IDs are retained in the machine-readable audit export.

D.3 Homogeneous MLX subset

For seeds 1–3, baseline-adjusted net gains are: answer–teacher gate +4.27 points (95% CI $[+0.53, +8.27]$, $p_{boot} = 0.034$, Holm $p = 0.204$); answer-only +3.20; medoid +2.93; gated U-OPSD +2.93; legacy ungated reference +2.67; vote weighting +2.67; reliability gating +1.87. Only the unadjusted bootstrap interval for answer–teacher gating excludes zero.

D.4 Complete five-seed secondary metrics

The final post-merge export preserves Single, Pass@8, invalid-output, and wrong-output changes for all 40 jobs. Table 13 reports baseline-adjusted five-seed means. For invalid and wrong rates, negative values are favorable. These quantities are secondary and descriptive; no additional multiplicity-corrected winner claim is made.

Variant	Single net Δ	Pass@8 net Δ	Invalid net Δ	Wrong Δ
Answer–teacher gate	+2.56	+1.60	-0.20	-1.60
Answer only	+2.08	+1.44	-0.12	-1.60
Medoid	+1.60	+1.92	-0.50	-1.44
Gated U-OPSD	+0.48	+0.96	-0.48	-0.96
Legacy ungated ref.	+0.96	-0.48	-0.20	-2.40
Vote weight	+1.92	-1.12	+0.38	-1.60
Reliability gated	+0.96	+0.80	-0.52	+0.32

Table 13: Complete five-seed secondary metrics from the final 40-job export, in percentage points. Net changes subtract the same seed's baseline re-measurement change; the baseline raw mean changes are -0.48 (Single), -0.16 (Pass@8), +0.28 (invalid rate), and 0.00 (wrong rate). Lower invalid/wrong rates are better.

The table clarifies why a single aggregate accuracy score is insufficient. Answer–teacher gating has the largest Single net gain, whereas medoid has the largest Pass@8 net gain and a larger reduction in invalid outputs. Vote weighting retains a positive Single gain but has negative Pass@8 net change and increased invalid output, driven especially by seed 2. The legacy ungated reference reduces wrong outputs most strongly but does not improve Pass@8 after baseline adjustment. These cross-metric reversals support reporting a vector of outcomes rather than ranking methods by one scalar.

E Post-Freeze U-OPSD Baseline-Fidelity Check

This secondary analysis was executed only after the original 40-job result file had been frozen. The original live and frozen result files remained byte-identical after the new runs (SHA-256 `48f31f594e9e...ae5f018`). The common M0 adapter was reloaded before each seed and verified at SHA-256 `01145f6bbe8b...cfe24`. The post-freeze cloud execution snapshot is identified by git `fb4d85b124cc...`. These checks prevent the secondary analysis from silently rewriting the confirmatory family.

Item	Post-freeze setting
Sampling / threshold	$G = 8$; self-consistency threshold $\tau = 0.5$
Pseudo-solution / teacher	shortest agreeing rollout
Student trajectory	one deterministically seeded uniform disagreeing rollout
Conditioning	teacher: problem + pseudo-solution + student prefix; student: problem + student prefix
Objective / teacher state	forward KL; fixed initial-policy teacher (`ema_decay=1.0`)
Reliability use	no RAST gate; no hidden answer or process gold used in training
Shared study stack	Qwen3-4B; HF/bf16/CUDA; seeds 4–5; temperature 0.7; max_tokens 2048; top-$k = 100$ KL; LoRA; one-pass training
Not an exact reproduction because	source-paper LoRA/training schedule is not reproduced; top-$k = 100$ approximation; no per-token pointwise KL clipping; shared temperature/token cap and one-pass optimizer are retained

Table 14: Protocol for the paper-structure-faithful post-freeze baseline. It is deliberately outside the original confirmatory family.

Seed	Samples	Tokens	Loss	Single Δ	Majority@8 Δ	Pass@8 Δ
4	26	6,765	0.445962	+0.8	+0.8	-0.8
5	27	10,829	0.948361	+2.4	+2.4	+3.2
Mean	–	–	–	+1.6	+1.6	+1.2

Table 15: Raw post-freeze changes in percentage points. No inferential test is attached because the check is post-freeze and contains only two training replications.

Seed	Invalid Δ	Wrong Δ	Frozen baseline Majority@8 Δ	Descriptive net Majority@8 Δ
4	+0.1	0.0	-0.8	+1.6
5	+0.4	-2.4	+3.2	-0.8
Mean	+0.25	-1.2	+1.2	+0.4

Table 16: Secondary and descriptive normalization for the post-freeze check, in percentage points. Lower invalid/wrong rates are better. The baseline adjustment reuses already-frozen same-seed re-measurements and is descriptive, not part of a pre-specified inferential family.

Across the two post-freeze seeds, raw Majority@8 changes are non-negative, but the descriptive baseline-adjusted mean is only +0.4 points and one seed is negative after adjustment. The two-seed result is of the same order as re-measurement noise, so it supports baseline fidelity without establishing equivalence, superiority, or a clean secondary-metric dominance pattern.

F Post-Freeze MATH-500 Public-Benchmark Downstream Stress Test

This study was specified and executed only after the original 40-job analysis was frozen. It uses the public 500-item MATH-500 evaluation split, derived from MATH/PRM800K (Hendrycks et al., 2021; Lightman et al., 2023), and a deterministic 256-item MATH training subset with zero normalized-problem overlap against MATH-500. Training receives only prompts/public metadata; reference solutions, final answers, and process labels are not supplied to collection or distillation. All three planned seeds (6–8) are reported.

Item	Frozen setting
Model / stack	Qwen3-4B; Hugging Face bf16/CUDA; LoRA on the shared study stack
Evaluation	MATH-500, 500 tasks; Math-Verify symbolic equivalence; ties count as abstention/incorrect
Collection pool	256 deterministically selected MATH training prompts; zero normalized-problem overlap with evaluation
Sampling	$G = 8$; temperature 0.7; max 2048 generated tokens
Optimization	forward KL; learning rate 10^{-6}; top-$k = 100$ approximation; batch size 2; one pass
U-OPSD condition	$\tau = 0.5$ eligibility; shortest agreeing teacher/pseudo-solution; one deterministic pseudo-uniform disagreement; fixed initial teacher; no RAST gate
RAST condition	answer–teacher reliability signal; v1.4 gate/control; reliability weighting; no hidden MATH gold/process annotations in training
Uncertainty	10,000 paired crossed bootstrap replicates; resample 3 seeds and 500 task IDs independently with replacement; preserve condition pairing
Inferential status	post-freeze secondary downstream stress-test evidence; outside the original Holm multiplicity family

Table 17: Frozen MATH-500 downstream stress-test protocol.

Seed	Condition	Samples	Task cov.	Single	Majority@8	Pass@8	Tie
6	M0	–	–	80.2	87.0	93.2	3.6
6	U-OPSD	63	24.6	80.2	85.4	93.2	5.6
6	RAST	68	14.8	81.6	85.6	92.6	6.2
7	M0	–	–	81.4	86.6	93.8	4.2
7	U-OPSD	65	25.4	81.8	86.6	92.2	5.0
7	RAST	78	16.4	80.6	85.8	92.2	3.6
8	M0	–	–	82.6	86.8	93.0	3.4
8	U-OPSD	67	26.2	81.6	85.4	93.2	6.0
8	RAST	78	15.2	83.6	85.8	93.6	5.8
Mean	M0	–	–	81.40	86.80	93.33	3.73
Mean	U-OPSD	65.0	25.39	81.20	85.80	92.87	5.53
Mean	RAST	74.7	15.49	81.93	85.73	92.80	5.20

Table 18: MATH-500 seedwise and mean results. Accuracy, coverage, and tie values are percentages. Invalid rate is 0 for every seed and condition. "Samples" counts distillation trajectories, whereas task coverage counts distinct retained collection tasks; RAST can emit multiple minority trajectories per retained task.

Contrast	Single Δ	Majority@8 Δ	Pass@8 Δ
U-OPSD − M0	-0.20 $[-2.20, +1.80]$	-1.00 $[-2.47, +0.60]$	-0.47 $[-2.07, +1.20]$
RAST − M0	+0.53 $[-1.53, +2.60]$	-1.07 $[-2.53, +0.40]$	-0.53 $[-2.13, +1.07]$
RAST − U-OPSD	+0.73 $[-1.67, +3.13]$	-0.07 $[-1.40, +1.33]$	-0.07 $[-1.40, +1.27]$

Table 19: Post-freeze MATH-500 contrasts in percentage points with paired two-way bootstrap 95% percentile intervals. These intervals display uncertainty and are not added to the original Holm family.

The frozen tie-rate contrasts (pp) are +1.80 $[0.00, +3.60]$ for U-OPSD–M0, +1.47 $[-0.80, +3.60]$ for RAST–M0, and -0.33 $[-2.20, +1.60]$ for RAST–U-OPSD. Tie-as-abstention was specified before evaluation, so tied cases are not favorably re-scored. As a separate descriptive sensitivity, conditioning on non-tied task–seed cells gives 90.17% for M0, 90.83% for U-OPSD, and 90.44% for RAST (Table 20). This changes the estimand from "all tasks" to "decided tasks" and cannot replace the pre-specified Majority@8 comparison; it shows only that the small all-task differences are entangled with condition-dependent tie frequency.

Condition	Correct majorities	Ties	Non-tied cells	Majority among non-ties
M0	1302	56	1444	90.17%
U-OPSD	1287	83	1417	90.83%
RAST	1286	78	1422	90.44%

Table 20: Post-hoc descriptive tie-excluded sensitivity pooled over three seeds (1,500 task–seed cells per condition). Tied cases are excluded rather than favorably broken. The changed denominator makes this a diagnostic of the tie mechanism, not an alternative primary accuracy result.

Task coverage is lower for RAST than for U-OPSD (15.49% vs. 25.39% on average), yet RAST emits more distillation trajectories (74.7 vs. 65.0 per seed) because the frozen RAST policy may retain multiple minority trajectories from an admitted task whereas the paper-structure-faithful U-OPSD condition uses exactly one deterministic disagreement trajectory. This difference prevents a simple "less supervision" interpretation and is why both coverage and sample volume are reported. The public-benchmark stress test does not show a Majority@8 gain for either learned condition; its contribution is to bound the downstream-transfer claim rather than to create a post-hoc winner.

Provenance. The MATH integration code, common M0, 256-item collection pool, 500 evaluation prompts/gold, data manifest, and pre-bootstrap analysis plan are all SHA-256 frozen. Full hashes, seedwise metrics, adapter hashes, and bootstrap summaries accompany the anonymous supplement; the manuscript prints only abbreviated identifiers where needed.

G Audit Trail and Reproducibility

Before final analysis, 56 historical runs were invalidated and retained. Table 21 gives a sanitized category-level ledger and pre-result validity rationale; this submission snapshot does not fabricate per-run identifiers that were not preserved in its export.

Category	Count	Exclusion rule / rationale
Degenerate sampling	31	Failed the pre-analysis sampling diversity/validity guard; the generated set could not support the planned comparison.
Zero-adapter / no-op	7	No effective adapter update; treated as technical no-op failures rather than negative scientific outcomes.
Automatic validity exclusions	18	Predefined execution/provenance failures including interrupted jobs, mock contamination, or baseline inconsistency; not selected by held-out accuracy.
Total invalidated	56	Retained in audit history; final inferential denominator is the separately re-measured 40/40 valid jobs.

Table 21: Sanitized invalid-run ledger. Exclusion status is based on execution/provenance validity rather than effect direction.

Before final analysis, 56 historical runs were invalidated and retained: 31 degenerate-sampling runs, seven zero-adapter no-op runs, and 18 automatic exclusions (including interrupted cloud jobs, one mock-contaminated run, and one inconsistent baseline). Auxiliary development/process-audit rounds remain in the archive but are not final-study denominators. The final weight table contains only re-measured/validated jobs; the common M0 SHA-256 prefix is `01145f6b...`, and the final state is 40 done, 40 valid, 0 invalid pending, `complete=True`. The post-freeze fidelity audit verifies shortest-teacher selection, one deterministic-uniform disagreement per task, no reliability gate, exact M0 reload, distinct non-M0 final adapters, and an unchanged original 40-job result artifact. The later MATH-500 downstream evaluation preserves the same original 40-job and M0 hashes, reports all three planned seeds, freezes all JSON/JSONL/adapters in a scientific-artifact manifest, and locks the bootstrap design before inspecting its confidence intervals.