WHAT CAN A SUCCESS-SET FLIP TELL US? INDEPENDENT RESPONSE VALIDATION IN LANGUAGE MODEL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We ask which claims about items selected by a finite-budget success-set flip survive responses withheld from selection, larger discovery budgets, and later checkpoints. Among 151 Llama export-comparison flips selected at eight attempts, independent V_{128} responses support 81 original directions; a separate W_{512} pool supports 96, favors 21 opposite directions, and leaves 34 unresolved. Nine reversals survive Fisher/Holm correction over both directions of all selected items, with a median signed effect of -9.18 percentage points. Across saved W request-index prefixes, corrected determinate directions rise from 56 at depth 128 to 91 at 512. Increasing the discovery budget to 64 makes 120 selected items solved at both endpoints, including 57 with independent directional support. A separate Qwen3-4B checkpoint study follows 95 items with independently confirmed improvements. Their fixed-denominator score declines by 12.25 points after t_1, with a model-conditional posterior interval of $[-12.94, -11.17]$, although 60 items retain support relative to baseline. An unmatched reference supplies descriptive context. Together, these fixed-item measurements connect a selected label to its later directional evidence under frozen machine scores. Training provenance and response dependence limit broader interpretation. The anonymous supplement provides analysis code and item-level results, which we will also release publicly.

1 INTRODUCTION

When two model exports are compared, an item that changes from no observed success to at least one can be labeled Learned, while the reverse change can be labeled Forgotten (Toneva et al., 2019a). These success-set labels make item-level results easy to inspect, and coverage across attempt budgets has become a useful view of on-policy distillation (OPD) evaluations (Chen et al., 2021; Agarwal et al., 2024a; Ge et al., 2026). Yet a label records an event in K sampled responses, not the item's success probability. For an item with success probability p, the chance of appearing solved is $1 - (1-p)^K$, which changes with K even if p does not. The stronger interpretation, that the probability changed in the labeled direction, needs evidence beyond the responses that produced the flip.

This distinction creates two concrete problems for an item selected by a small-budget flip. Selection favors response blocks whose observed outcomes differ sharply, even when the underlying difference is weak or points the other way. Increasing K can then make both endpoints appear solved, concealing a probability difference rather than resolving it. Repeated-response evaluation already offers estimates of success probabilities and their uncertainty (Hariri et al., 2026), and prior OPD work reports probability changes within Retained items (Ge et al., 2026). We therefore do not introduce a new estimator or claim that Retained implies unchanged performance. Instead, we ask what becomes of a *particular selected item* when new responses test its original direction, the discovery budget changes, or the model is measured at later checkpoints. The responses used for selection must be kept distinct from those used to answer that question. We first fix 151 Llama Native/GAD items identified as flips at $K = 8$ and measure them in independent response pools. At 128 responses per endpoint, V supports 81 original directions, favors nine opposites, and leaves 61 uncertain. A separate 512-response pool, W, supports 96 originals and 21 opposites, with 34 uncertain; nine of those opposite directions survive Fisher/Holm correction over both signs of all 151 items. The endpoint counts and effect intervals of those nine are reported individually, making the contrary evidence

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

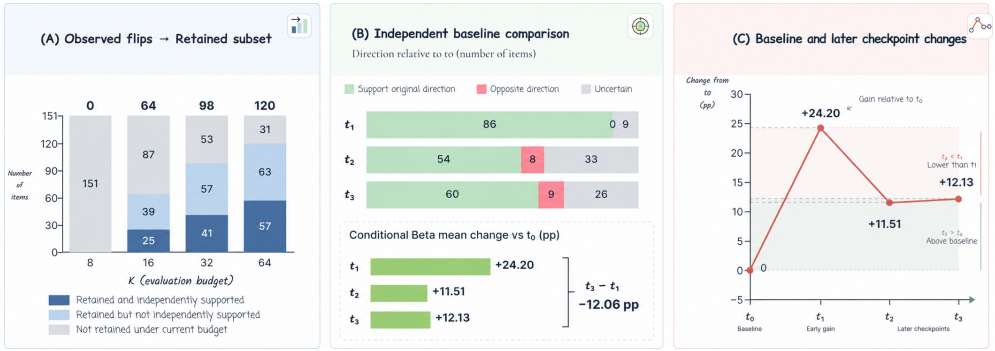

Figure 1: Two fixed-item studies. (A) Llama discovery labels as K grows, with independent V support among Retained items. (B) Baseline-relative V directions for 95 Qwen items. (C) Their conditional Beta mean changes from t_0.

inspectable. Saved request-index prefixes show that corrected determinate decisions increase from 56 at depth 128 to 91 at 512, although the prefixes share responses and cannot establish an optimal stopping budget. Meanwhile, extending the *discovery* budget to $K = 64$ makes 120 of the same items Retained, including 57 whose original direction still has independent V support (Figure 1A). Thus the direction and the success-set label answer different questions on this fixed cohort.

Our second study tracks improvement-selected items across one Qwen3-4B checkpoint sequence. A greedy screen finds 171 candidates; separate C_{128} responses confirm 95 increases in scored success probability, and a pool V tracks later outcomes. Their baseline conditional mean is already 54.76%, so confirmation reflects improvement, not emergence from zero ability. Although the group stays above baseline at t_3, its fixed-denominator V score drops from 79.24% at t_1 to 67.00% at t_3 (Figure 1B,C). A family-hash reference list provides context: after removing five shared items, R123 falls by 0.30 points, yielding a C95–R123 contrast of -11.95 points. Because the reference differs in selection history and baseline distribution, this contrast does not identify the effect of improvement selection or OPD training.

Together, the studies connect the response block that selects an item to separate measurements of its direction, budget sensitivity, and later checkpoint behavior. A selected-item conclusion should state the selection budget, held-out endpoint counts, directional uncertainty, and the checkpoint to which it applies. These fixed-item, fixed-export measurements cannot substitute for independent training runs and do not identify an OPD effect or a general law of capability change.

2 Measurement Design

The design follows three linked empirical questions (Figure 2). First, do directions measured with new responses agree with selected success-set flips? Second, what remains visible as the discovery budget grows on those same items? Third, after separate confirmation of an improvement, how do baseline-relative and t_1-relative outcomes compare at later checkpoints? The Llama Native/GAD pair supplies the direction and budget study. Four Qwen3-4B exports supply the checkpoint study and a post hoc reference comparison. Related-item analyses and completed controls are reported separately in the appendix. Appendix Table 7 maps every pool and cohort to its role and overlap.

For model state M, item x, response block b, and attempt budget K, the observed solved indicator is

$$S^{(b)}_{M,K}(x) = \mathbf{1}\left\{\sum_{j=1}^{K} s_x(y^{(b)}_{M,j}) > 0\right\}, \tag{1}$$

where s_x is the frozen binary scorer. Before/after pairs $(0,1)$, $(1,0)$, $(1,1)$, and $(0,0)$ receive the labels Learned, Forgotten, Retained, and Unsolved. The separate target $p_{M,\pi}(x)$ is the probability that a response receives a scored success under evaluation specification π, including prompt, de-

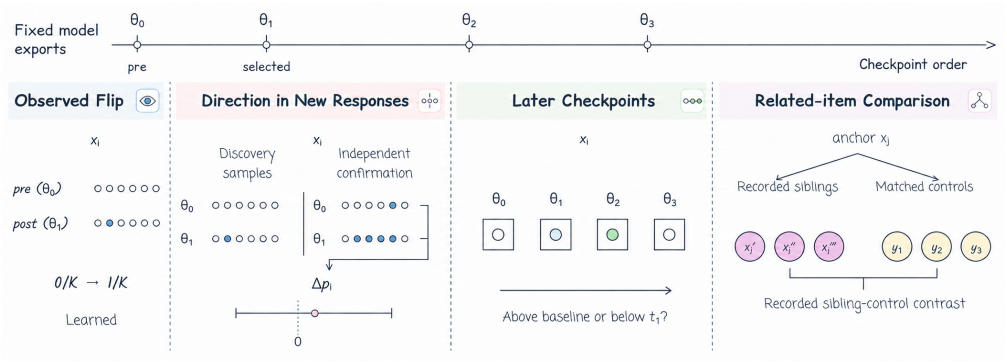

Figure 2: Selection and independent measurement in the Llama and Qwen studies. Arrows follow response pools and checkpoint order; related-item comparisons are descriptive.

coding, output limit, and scorer. It is an operational response probability, not a human-adjudicated probability of mathematical correctness. Unresolved scores remain explicit.

2.1 Select transitions and measure their directions

We use Llama-3.2-3B-Instruct as the Native endpoint and the archived GAD export as the comparison endpoint on IFEval and GSM8K. From historical pool D, with 128 scored responses per item and endpoint, a frozen hash ordering defines a $K = 8$ discovery prefix. All Learned and Forgotten items enter the fixed cohort. Separate pools V_{128} and W_{512} measure the original direction, without revising cohort membership. Table 6 records the recoverable model and inference configuration; the GAD training recipe is not established by the evaluation archive.

For c binary successes and w failures, the working model uses independent Bernoulli trials at a fixed endpoint probability and a Beta$(1, 1)$ prior, giving $p \mid c, w \sim \text{Beta}(1 + c, 1 + w)$. For a Learned candidate, we evaluate $P(p_{\text{after}} > p_{\text{before}} \mid \text{scores})$; for a Forgotten candidate, the inequality is reversed. Probabilities at least 0.975 support the original direction, those at most 0.025 support the opposite, and intermediate values are uncertain. Each pool is analyzed separately, and a joint item table records agreement and movement between their decisions.

Independent validation withholds responses from selection but does not establish an iid generation stream. We therefore add a post hoc calibration that uses the actual $K = 8$ selection, validation depths, Beta$(1, 1)$ prior, and 0.975 threshold. Exact count enumeration covers unchanged and changed probabilities, followed by beta-binomial batch sensitivities (Appendix I). On the observed data, we also apply Fisher/Holm over all 302 one-sided hypotheses, separately within each pool, and a stricter posterior threshold $1 - 0.05/302$. These are sensitivity analyses of the frozen cohort, without revising its membership. For a separate depth analysis, we recompute both rules on nested prefixes of saved W request indices while holding all 151 selected items fixed. These prefixes reuse responses and lack sequential stopping guarantees or response independence.

2.2 Extend the budget without reselecting items

To isolate how the label changes with budget, we retain the $K = 8$ cohort and extend its original D prefixes to $K = 16, 32, 64$. Each new label is paired with the unchanged V direction decision. Under an illustrative iid model, the probability of at least one success is

$$q_K(p) = 1 - (1 - p)^K. \tag{2}$$

Thus two endpoints can both become solved at a larger K even when their per-response probabilities differ. The empirical analysis asks which selected items exhibit this pattern. Nested D prefixes and subsets of W reuse responses and provide sensitivity views of fixed model states.

2.3 CONFIRM IMPROVEMENT AND REVISIT LATER CHECKPOINTS

The Qwen study evaluates four hash-bound Qwen3-4B exports, t_0 through t_3. Publisher cards describe an off-policy endpoint followed by cumulative exposures of 10k, 15k, and 20k on-policy prompts; these are exposure labels, not verified optimizer steps. A historical greedy screen of 4,991 fixed anchors selects t_0-incorrect/t_1-correct candidates. Separate C_{128} responses assess their t_0-to-t_1 direction with the rule above. Where scores are unresolved, confirmation requires support even when all missing t_0 outcomes are successes and all missing t_1 outcomes are failures. The confirmed item list is fixed before the subsequent V analysis. We call this confirmed improvement: the test concerns an increase in scored success probability and imposes no zero-probability baseline requirement.

We then measure baseline-relative directions at t_1, t_2, t_3 and directly pair each item's t_1 and t_3 responses. The scored-response Beta mean is $(c+1)/(c+w+2)$. For the group comparison, fixed-denominator scored success is c/n with $n = 128$; if m outcomes are unknown, completing them permits values in $[c/n, (c+m)/n]$. Conditional means and fixed-denominator yields are different estimands. Their missing-outcome ranges are deterministic bounds, not sampling intervals. For uncertainty in the frozen correct-yield target, a response is one only if its frozen score is correct. With group membership fixed, independent Beta$(c+1, n-c+1)$ endpoint draws yield posterior means and pointwise 95% equal-tailed intervals for each group change and their contrast. Unknown scores stay in the original non-correct remainder for this target; completing them defines a separate sensitivity analysis.

2.4 PLACE THE SELECTED CHANGE BESIDE A REFERENCE

R128 contains anchors from 128 result-independently hash-selected reference families. Its later V extension is post hoc and does not condition reference membership on the greedy-to-C confirmation criterion. Removing their five shared items gives the mutually exclusive R123 comparison. We complete 128 responses per endpoint using 237 newly sampled item-checkpoint cells and 19 pre-existing V cells, then subtract the R123 mean $t_3 - t_1$ change from the C95 change. The extension supplies a descriptive reference for these fixed groups; it does not balance initial success probabilities, task composition, or selection history. A common-S measurement remains a separate sensitivity check. To examine covariate overlap, a post hoc match uses only S baseline scores, early gains, and task domains, then compares the matched pairs' separate V outcomes. This adjustment does not reconstruct the greedy-to-C reference selector.

Related-item comparisons and separate model settings use their own populations and rules. Their definitions and completed results appear in Appendices D–H; they are not pooled as replications of either main study.

3 EXPERIMENTS

The main Llama study uses 541 IFEval and 1,319 GSM8K items, with 151 transitions selected from D at $K = 8$. IFEval uses strict prompt-level instruction satisfaction; GSM8K uses the frozen numeric extraction score. The Qwen3-4B study uses UGMathBench anchors and its frozen answer-slot scorer. Llama sampling uses temperature 0.8, top-$p = 0.95$, and a 1,536-token output cap; Qwen C/V uses 0.7, 0.95, and 32,768 tokens. Table 6 supplies further configuration and provenance. Unless specified otherwise, directional decisions use the itemwise Beta$(1, 1)$ threshold 0.975. All counts concern fixed items and exported weights; response pools do not constitute independent training runs.

3.1 INDEPENDENT RESPONSES CHANGE SELECTED-ITEM INTERPRETATIONS

We first test whether selected flips predict a direction in responses withheld from selection. Table 1 compares the original decisions with stricter rules on the same items. The historical $K = 1$ through 64 A/B record is retained in Appendix Table 8; its partitions differ from the main $D/V/W$ analysis.

Pool	Direction rule	Original	Opposite	Uncertain
V_{128}	Beta$(1,1)$, 0.975	81	9	61
	Beta$(1,1)$, $1-0.05/302$	57	0	94
	Fisher/Holm, 302 hypotheses	54	0	97
W_{512}	Beta$(1,1)$, 0.975	96	21	34
	Beta$(1,1)$, $1-0.05/302$	82	9	60
	Fisher/Holm, 302 hypotheses	82	9	60

Table 1: Direction decisions for all 151 fixed Llama candidates in independent V_{128} and W_{512} pools. Holm corrects both one-sided tests per item across 302 hypotheses; posterior rows use the stated Beta$(1,1)$ thresholds.

Independent V_{128} responses reveal a substantial task difference within the fixed cohort. They support 59 of 75 IFEval directions, favor two opposite directions, and leave 14 uncertain. For the 76 GSM8K candidates, the corresponding counts are 22, seven, and 47 (Table 3). The difference appears within both the originally Learned and Forgotten strata. An uncertain direction means that the rule does not resolve the comparison at this depth; it does not imply equal endpoint probabilities.

The deeper W_{512} pool shows that additional measurement can support the opposite of the discovery label. It supports 96 original directions, favors 21 opposite directions, and leaves 34 uncertain. Of the 21 opposite decisions, seven are on IFEval and 14 on GSM8K. Their observed endpoint changes, signed toward the original discovery direction, have a median of -5.66 percentage points and range from -17.77 to -1.95. Nine reversals survive both stricter rules in Table 1, with signed effects from -17.77 to -6.05 points and median -9.18. Among those nine, five also have posterior probability at least 0.975 of a reversal exceeding five points. This last magnitude check is pointwise and uses a post hoc threshold; it is not an additional multiplicity-controlled claim. Table 2 exposes the nine corrected reversals. IFEval 2941 changes from zero versus one discovery success to 123 versus 32 W successes. GSM8K 1145 has the opposite discovery label, yet its W correct counts increase from 27 to 74. In both cases, the held-out direction contradicts the label. The displayed intervals are pointwise posterior summaries after item selection, without simultaneous or selection-adjusted coverage.

Item	D	Native	GAD	Δ_D (pp)	Holm p
IFEval 2785	L	42	11	$-6.05_{[-8.84,\,-3.38]}$	1.50×10^{-3}
IFEval 2941	L	123	32	$-17.77_{[-22.00,\,-13.48]}$	9.65×10^{-14}
IFEval 3327	F	21	62	$-8.01_{[-11.36,\,-4.70]}$	3.41×10^{-4}
IFEval 3362	L	78	37	$-8.01_{[-11.86,\,-4.15]}$	6.93×10^{-3}
IFEval 3456	L	39	7	$-6.25_{[-8.87,\,-3.77]}$	1.22×10^{-4}
GSM8K 1094	F	56	109	$-10.35_{[-14.79,\,-5.86]}$	9.39×10^{-4}
GSM8K 1129	L	139	73	$-12.89_{[-17.74,\,-7.94]}$	5.13×10^{-5}
GSM8K 1145	F	27	74	$-9.18_{[-12.82,\,-5.56]}$	1.08×10^{-4}
GSM8K 1174	F	20	83	$-12.30_{[-15.94,\,-8.69]}$	3.74×10^{-9}

Table 2: Nine Llama W_{512} reversals with Fisher/Holm support across 302 hypotheses. Native and GAD give correct counts out of 512; L/F are discovery labels. Negative Δ_D opposes that label. Gray subscripts are pointwise 95% Beta$(1,1)$ credible intervals in pp.

Task	D label	Candidates	V supported	V opposite	V uncertain
IFEval	Learned	21	13	2	6
IFEval	Forgotten	54	46	0	8
GSM8K	Learned	41	15	2	24
GSM8K	Forgotten	35	7	5	23
Same-item V decision		Items	W supported	W opposite	W uncertain
Supported		81	79	1	1
Opposite		9	0	7	2
Uncertain		61	17	13	31

Table 3: Independent direction decisions for the 151 D-selected Llama items. Upper rows split V_{128} by task and discovery label; lower rows pair each item with its separate W_{512} decision. Directions use Beta$(1, 1)$ at 0.975 relative to D.

The joint table distinguishes uncertainty from contrary evidence: among the 61 items uncertain in V, W supports 17 original and 13 opposite directions. Five of the nine corrected W reversals have nominal opposite support in V; four are uncertain. All nine have contrary observed signs in each of four hash partitions and in each of four request-index segments. Six survive every segment omission with 302-hypothesis correction under either partition scheme. These overlapping subsets are post hoc checks of one pool, not independent runs (Appendix I.4). Figure 3 gives the nested-depth view.

Validation depth and directional yield. On nested W request-index prefixes, Fisher/Holm resolves 56, 80, 89, and 91 of the 151 selected items at depths 128, 256, 384, and 512; the opposite-direction counts are zero, five, eight, and nine (Appendix Table 49). All five corrected opposites at 256 and all eight at 384 belong to the final nine. These prefixes share responses and describe decision yield within one cohort. Figure 3 instead uses the archived block order for its taskwise 128- and 256-response points.

Matching calibration exposes sensitivity to response dependence. On the artificial no-change grid, nominal direction rates are 3.88% at depth 128 and 4.77% at 512; correlation 0.01 within batches of 128 raises them to 16.24% and 18.62%. The data's actual correlation is unknown. Scoring sensitivity is also item-specific: changing five to 46 adverse binary labels can remove a given reversal's Holm support, although erasing its observed sign requires 31 to 91 changes (Appendix Table 42). Neither check provides human validation of the frozen scores.

3.2 A LARGER BUDGET CONCEALS SOME INDEPENDENTLY SUPPORTED DIFFERENCES

We next keep the selected item list and every V decision fixed, extending only the discovery prefixes. At $K = 8$, all 151 candidates are flips by construction. At $K = 16$, 64 are Retained, increasing to 98 at $K = 32$ and 120 at $K = 64$. Within these Retained subsets, 25, 41, and 57 items still have V support for their original direction (Figure 1A). The change concerns the label used to display the same comparison, rather than a disappearance of its independently measured direction.

At $K = 64$, the taskwise counts are 51 of 75 Retained items on IFEval and 69 of 76 on GSM8K, including 37 and 20 with V support (Figure 4). Equation 2 explains the general tendency of any-success labels to merge, while these counts quantify it in the fixed cohort. They neither extrapolate to other discovery samples nor reproduce a $K = 1024$ capability-boundary analysis.

3.3 BASELINE SUPPORT CAN COEXIST WITH A LATER DECLINE

The Qwen3-4B study asks whether independently confirmed improvements persist across later checkpoints. The greedy screen yields 171 candidates from 4,991 fixed anchors. Separate C_{128} responses confirm 95: 79 have complete binary scores and 16 also pass the least-favourable missing-outcome completion. These C95 items define the cohort for later analysis; their number is conditional on the greedy screen and does not estimate the prevalence of zero-to-positive capability transitions.

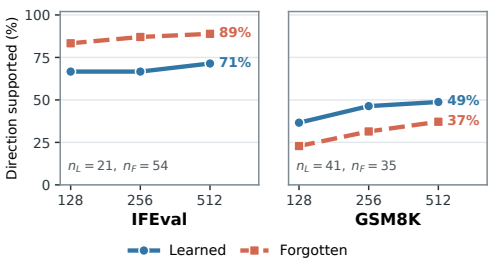

Figure 3: Independent W support for the original direction in fixed Llama $K = 8$ items. Blue solid lines denote Learned and coral dashed lines Forgotten; response depths use archived block order.

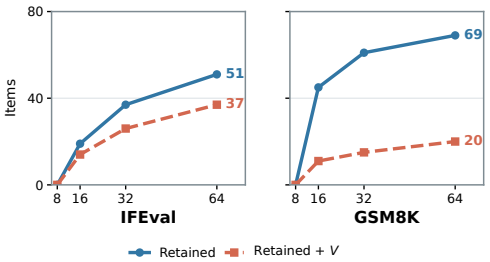

Figure 4: Fixed Llama $K = 8$ items that become Retained as discovery K grows. Dashed lines count those with independent V_{128} support; panels show IFEval ($n = 75$) and GSM8K ($n = 76$).

Relative to t_0, separate V responses support 86 C95 items at t_1, 54 at t_2, and 60 at t_3 (Figure 1B). At t_2, eight items have opposite and 33 uncertain directions; at t_3, the counts are nine and 26. Following the same 86 items supported at t_1, 57 remain supported at t_3, 23 become uncertain, and six receive opposite support. These baseline-relative outcomes leave open whether the selected items maintain their t_1 response rates.

Direct time comparisons answer that question. The C95 item-equal conditional Beta mean is 54.76% at t_0, 78.95% at t_1, 66.27% at t_2, and 66.89% at t_3. The final checkpoint remains 12.13 points above baseline but 12.06 points below t_1. All 171 candidates have a conditional change of -7.79 points. Appendix Table 17 separately reports fixed-denominator changes and unknown-score bounds. Table 4 gives the C95 checkpoint means and direction counts. Moreover, its empirical pass@128 is 100% at both t_1 and t_3, despite the per-response decline (Appendix D.1).

Checkpoint	Beta mean (%)	Change vs. t_0 (pp)	Change vs. t_1 (pp)	Direction vs. t_0: S/U/O
t_0	54.76	0.00	-24.20	self
t_1	78.95	$+24.20$	0.00	86/9/0
t_2	66.27	$+11.51$	-12.69	54/33/8
t_3	66.89	$+12.13$	-12.06	60/26/9

Table 4: Independent V results for the same 95 Qwen items across t_0 to t_3. Means are item-equal conditional Beta means; S/U/O count baseline-relative supported, uncertain, and opposite directions.

3.4 An unmatched reference gives descriptive context

Before comparing changes, we distinguish the groups' selection histories. C95 membership requires greedy-to-C confirmation; R128 membership depends only on family hash. Its endpoint V extension is a post hoc measurement of a fixed reference list. Replaying the historical selector on R128 yields eight candidates and five confirmed items, all already in C95. Removing those five gives R123. The mutually exclusive groups are therefore not matched for confirmed improvement, initial success rate, or task composition. Even their empirical pass@128 coverage differs: C95 is 100% at both endpoints, while R123 changes from 68.29% to 67.48%.

With 128 responses per item and endpoint, the frozen fixed-denominator score for C95 falls from 79.24% to 67.00%, or -12.25 percentage points. R128 moves from 57.58% to 57.32%, or -0.26 points; R123 moves from 56.19% to 55.89%, or -0.30 points. Subtracting the R123 change gives -11.95 points. This records different mean trajectories for the selected and reference items. A near-zero reference estimate is not an equivalence result, and the contrast cannot establish that improvement-selected items are generally less stable.

Table 5 directly reports uncertainty in the fixed-group correct-yield changes. C95's 95% posterior interval is $[-12.94, -11.17]$ points and the contrast interval is $[-12.73, -10.80]$; R123's interval crosses zero. These summaries condition on independent responses and fixed item membership. They neither quantify training-run variation nor correct the groups' different selection histories.

Fixed group or contrast	Empirical change	Posterior mean	95% CrI
C95	-12.25	-12.06	$[-12.94, -11.17]$
R123	-0.30	-0.29	$[-0.69, +0.10]$
C95 minus R123	-11.95	-11.76	$[-12.73, -10.80]$

Table 5: Qwen V changes from t_1 to t_3 in fixed-denominator correct yield. Intervals use independent Beta$(c+1, 129-c)$ endpoint draws with fixed item membership; unknown scores retain their frozen treatment.

The separate V responses matter because the decline is measured without reusing the C block that admitted an item to C95. That separation addresses same-response selection noise, while leaving selection-induced differences in the underlying item populations intact. A fixed-item no-change simulation places the C95 decline beyond its iid response-sampling null (Appendix Table 31); it does not model the whole selection process or supply an OPD control. The recovered common-S responses give a C95 decline of 11.55 points, with 63 of 95 items declining in both pools. Every actual S worker has a negative mean change (Appendix G.1).

Identified unresolved V scores cannot reverse the direction: adverse completions give C95 bounds $[-12.43, -12.07]$ points and contrast bounds $[-13.42, -10.86]$. These are deterministic bounds, separate from sampling uncertainty and undiscovered semantic errors. The decline is not confined to the ten largest contributors. C95 contains 68 declining, six tied, and 21 increasing items; removing those ten leaves a -7.10-point mean change among 85 items. Removing any one recorded domain leaves a change between -12.71 and -11.35 points (Appendix Tables 40 and 41). These post hoc checks measure concentration, not scorer validity.

Composition remains consequential. Matching on separate S baseline scores, early gains, and domain retains five, 16, or 35 pairs under five-, ten-, or 20-point calipers. The respective V contrasts are -9.38, -2.73, and -5.74 points (Appendix Table 36). Hypothetical AI-label adjustments also keep the contrast negative under their stated assumptions (Appendix G). Within the same 171 greedy-selected candidates, the 76 not confirmed by C decline by 2.35 points in V, giving a C95 minus U76 difference of -9.89 (Appendix Table 35). Their confirmation histories still differ; none of these comparisons supplies a matched experimental control.

4 Discussion and Limitations

The two studies distinguish a selected label from a direction measured in responses withheld from selection. Nine Llama items have corrected contrary evidence, while larger discovery budgets make some independently supported differences invisible to success-set labels. Nested W prefixes show increasing corrected decision yield on this cohort, but do not establish a sufficient validation budget or optimal response allocation. An item-level report should therefore give its selection budget, endpoint counts, directional rule, and uncertainty in the score difference.

The Qwen items improve in scored success probability from a substantial baseline, remain above that baseline on average, and then decline after t_1. The reference groups have different selection histories, and matching retains few items with varying restricted contrasts. One checkpoint sequence does not establish a general law of acquisition stability; related-item and other-setting analyses remain in the appendix.

Response independence and semantic scorer validity remain unverified. Synthetic calibration specifies the cost of dependence, while recorded partitions show consistency for some reversals without identifying physical batches. Individual adjusted reversals lose support after five to 46 adverse binary-score changes. The targeted packet covers 52 primary-V and 2,560 corresponding S answers, without completed human adjudication. Separate MATH-500 and D2 cohorts yield no supported reversals under the main rules (Appendix Table 24).

Finally, hashes identify exports more completely than training histories. The GAD recipe remains unrecovered; Qwen publisher cards describe its sequence without a complete optimizer record. We consequently report fixed-model measurements motivated by OPD, not a systematic OPD training effect or replicated mechanism.

5 Related Work

Distillation and success sets. Knowledge distillation transfers teacher information through output distributions or generated sequences (Hinton et al., 2015; Kim & Rush, 2016). On-policy methods instead train on student-sampled trajectories (Agarwal et al., 2024b; Gu et al., 2023); recent analyses examine when that supervision helps and how its effects depend on the training setting (Li et al., 2026; Armandpour et al., 2026; Fu et al., 2026; Ding & Zhang, 2026). Closest to our use of success sets, Ge et al. (2026) studies OPD across test-time budgets and reports accuracy gains even within Retained items. We do not treat Retained as constant success probability. Our empirical target is different: the later evidence for the particular items chosen by a small-budget flip, with discovery responses kept separate from confirmation responses.

Repeated-response evaluation. Pass@K summarizes whether repeated sampling finds any successful answer (Chen et al., 2021), while studies of inference compute and reasoning stability examine how coverage and response behavior change with budget (Brown et al., 2024; Snell et al., 2024; Dragoi et al., 2025; Liu et al., 2025). Hariri et al. (2026) develops posterior estimates and uncertainty for success probabilities; other work studies efficient response allocation and statistical guarantees (Wu et al., 2026; Zhou et al., 2026; Kim et al., 2026). We use standard Beta and Fisher tests rather than a new estimator. The distinction here is between estimating performance on a fixed item and testing a claim made after that item was selected by an observed flip. Separate V/W pools address reuse of discovery responses, while multiplicity and dependence analyses describe the remaining limits of directional decisions (Dror et al., 2018; Bouthillier et al., 2021).

What an evaluation change measures. Scores can vary with prompts and execution conditions (Sclar et al., 2024; Bjarnason et al., 2026), and benchmark validity depends on whether the measured behavior matches the intended construct (Bean et al., 2025; Ribeiro et al., 2020). Evaluation protocol and training-task alignment can also alter broad model comparisons (Laban et al., 2025; Dominguez-Olmedo et al., 2025). Separately, training trajectories can show a decline in one measured behavior while other benchmark scores continue to rise (Goyal et al., 2025; Toneva et al., 2019b). These results motivate careful interpretation, but none makes the selected-item response direction a training effect. Our Llama and Qwen analyses therefore report fixed-model, fixed-item measurements, with scoring and provenance limitations kept explicit.

6 Conclusion

We measured what selected success-set flips mean when the same items are evaluated with responses withheld from selection. The Llama study finds both persistent original directions and nine reversals that also survive stricter statistical checks, while larger discovery budgets can hide differences that remain supported. On one Qwen3-4B trajectory, items with separately confirmed improvements remain above baseline on average despite a later decline, with a smaller descriptive change in an unmatched reference group. For claims about a selected item, these results motivate reporting the selection budget together with the later endpoint counts, directional uncertainty, and the checkpoint or budget to which the conclusion applies.

Reproducibility Statement

The accompanying anonymous supplement contains frozen counts, nine-item evidence, response-partition checks, available model bindings, and code for calibration, uncertainty, and matching analyses. Its README gives the reproduction command; Appendix J maps results to files. Local usernames and machine paths are removed. The recovered S endpoint texts accompany their score links; primary-V/W raw-text coverage and training histories remain incomplete. We will release these materials publicly.

AI Use Statement

Generative AI assisted with writing, code, evidence review, and figures.The authors are responsible for all content.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Piotr Stanczyk, Sabela Ramos, Sylvain Gelly, and Olivier Pietquin. On-policy distillation of language models: Learning from self-generated data. *arXiv preprint arXiv:2406.12009*, 2024a.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *International Conference on Learning Representations*, 2024b. URL https://arxiv.org/abs/2306.13649.

Mohammadreza Armandpour, Fatih Ilhan, David Harrison, Ajay Jaiswal, Duc N. M Hoang, Fartash Faghri, Yizhe Zhang, Minsik Cho, and Mehrdad Farajtabar. Unmasking On-Policy Distillation: Where It Helps, Where It Hurts, and Why. *arXiv preprint arXiv:2605.10889*, 2026. URL https://arxiv.org/abs/2605.10889v2.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Kaili-May Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott A. Hale, Inioluwa Deborah Raji, Christopher Summerfield, Philip H. S. Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what Matters: Construct Validity in Large Language Model Benchmarks. *arXiv preprint arXiv:2511.04703*, 2025.

Bjarni Haukur Bjarnason, André Silva, and Martin Monperrus. On Randomness in Agentic Evals. In *ICLR 2026 Workshop on Agents in the Wild*, 2026. URL https://arxiv.org/abs/2602.07150v3.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for Variance in Machine Learning Benchmarks. *arXiv preprint arXiv:2103.03098*, 2021. URL https://arxiv.org/abs/2103.03098.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large Language Monkeys: Scaling Inference Compute with Repeated Sampling. *arXiv preprint arXiv:2407.21787*, 2024. URL https://arxiv.org/abs/2407.21787.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Yi Ding and Ruqi Zhang. Does On-Policy Distillation Really Distill? From Noisy Teacher to Self-Improvement. *arXiv preprint arXiv:2608.31046*, 2026. URL https://arxiv.org/abs/2608.31046v1.

Ricardo Dominguez-Olmedo, Florian E. Dorner, and Moritz Hardt. Training on the Test Task Confounds Evaluation and Emergence. *arXiv preprint arXiv:2407.07890*, 2025. URL https://arxiv.org/abs/2407.07890v3.

Marius Dragoi, Ioana Pintilie, Florin Gogianu, and Florin Brad. Beyond Pass@k: Breadth-Depth Metrics for Reasoning Boundaries. *arXiv preprint arXiv:2510.08325*, 2025. URL https://arxiv.org/abs/2510.08325v2.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, 2018. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting On-Policy Distillation: Empirical Failure Modes and Simple Fixes. *arXiv preprint arXiv:2603.25562*, 2026. URL https://arxiv.org/abs/2603.25562v2.

Xinmu Ge, Zizhuo Zhang, Yu Huang, Jianing Zhu, Lin Yuan, Wanli Gu, Weichang Wu, Weiran Huang, Xiaolu Zhang, Bo Han, Jun Zhou, and Jiangchao Yao. Towards Understanding On-Policy Distillation through the Lens of Test-Time Scaling. *arXiv preprint arXiv:2608.11829*, 2026. URL https://arxiv.org/abs/2608.11829.

Sachin Goyal, Christina Baek, J. Zico Kolter, and Aditi Raghunathan. Context-Parametric Inversion: Why Instruction Finetuning Can Worsen Context Reliance. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10796v3.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: On-Policy Distillation of Large Language Models. *arXiv preprint arXiv:2306.08543*, 2023. URL https://arxiv.org/abs/2306.08543.

Mohsen Hariri, Amirhossein Samandar, Michael Hinczewski, and Vipin Chaudhary. Don't Pass@k: A Bayesian Framework for Large Language Model Evaluation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.04265v4.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Eungyeup Kim, Chenchen Gu, Vashisth Tiwari, and J. Zico Kolter. Measuring Five-Nines Reliability: Sample-Efficient LLM Evaluation in Saturated Benchmarks. *arXiv preprint arXiv:2605.11209*, 2026. URL https://arxiv.org/abs/2605.11209v1.

Yoon Kim and Alexander M. Rush. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016. URL https://aclanthology.org/D16-1139/.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs Get Lost In Multi-Turn Conversation. *arXiv preprint arXiv:2505.06120*, 2025. URL https://arxiv.org/abs/2505.06120v1.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking On-Policy Distillation of Large Language Models: Phenomenology, Mechanism, and Recipe. *arXiv preprint arXiv:2604.13016*, 2026. URL https://arxiv.org/abs/2604.13016.

Junnan Liu, Hongwei Liu, Linchen Xiao, Ziyi Wang, Kuikun Liu, Songyang Gao, Wenwei Zhang, Songyang Zhang, and Kai Chen. Are Your LLMs Capable of Stable Reasoning? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17594–17632, 2025. doi: 10.18653/v1/2025.findings-acl.905. URL https://aclanthology.org/2025.findings-acl.905/.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11324.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. *arXiv preprint arXiv:2408.03314*, 2024. URL https://arxiv.org/abs/2408.03314.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical investigation of catastrophic forgetting in gradient-based neural networks. In *International Conference on Learning Representations (ICLR)*, 2019a. URL https://openreview.net/forum?id=BJl6NsC-KS.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An Empirical Study of Example Forgetting during Deep Neural Network Learning. In *International Conference on Learning Representations*, 2019b. URL https://arxiv.org/abs/1812.05159.

Skyler Wu, Yash Nair, and Emmanuel J. Candès. Efficient Evaluation of LLM Performance with Statistical Guarantees. *arXiv preprint arXiv:2601.20251*, 2026. URL https://arxiv.org/abs/2601.20251v3.

Zhijian Zhou, Zesheng Ye, Zhaorun Chen, Bo Li, and Feng Liu. CELEUS: Certifiable and Efficient LLM Evaluation via E-Processes. *arXiv preprint arXiv:2606.20820*, 2026. URL https://arxiv.org/abs/2606.20820v2.

A Evaluation Configuration and Provenance

Table 6 distinguishes the evaluated artifacts from claims about how they were trained. Llama Native/GAD is a fixed exported-model comparison. The protocol identifies the Native tokenizer and configuration as Llama-3.2-3B-Instruct and binds both endpoints by hashes, but the recovered package does not establish the expanded meaning or executed training recipe of the GAD label. We retain that label for correspondence with the source records rather than inventing missing training details.

Field	Llama direction and budget study	Qwen confirmation and checkpoint study
Evaluated weights	Llama-3.2-3B-Instruct Native; hash-bound GAD export, using the Native tokenizer.	Four distinct Qwen3-4B exports. Cards identify the student as Qwen3-4B-Instruct-2507.
Training metadata	GAD is the retained archive label. Its expanded name, teacher, training data, and executed optimizer recipe are not recovered.	Cards name Qwen3-30B-A3B-Instruct-2507 as teacher and OpenThoughts-3 math slices as training prompts; execution lineage is incomplete.
Evaluation data	IFEval 541 and GSM8K 1,319; $K = 8$ selects 151 items. MMLU-Pro 2,000 is secondary.	UGMathBench: 4,991 anchors in the greedy screen, 171 candidates, C95, and the fixed R128 reference.
Response roles	D: 128 per endpoint before prefix selection; V: 128; W: 512 for each selected item and endpoint.	Greedy screen: one response per item; C/V: 128 per endpoint. R128 extends V at t_1, t_3 after the original analysis.
Sampling	Temperature 0.8, top-$p = 0.95$, BF16. W records unrestricted top-k and request-level seeds.	C/V: temperature 0.7, top-$p = 0.95$, unrestricted top-k, BF16. Greedy uses temperature zero.
Output / context	1,536 generated tokens / 4,096 total tokens. W has 2,276 length-limit finishes and no recorded unknown scores.	32,768 generated tokens / 40,960 total tokens. Length limits and scoring failures are tracked separately.
Prompt and runtime	Shared Native chat template; item-level rendered prompt hashes. W: vLLM 0.17.1, Transformers 5.2.0.	Frozen UGMathBench prompt builder; empty system message; checkpoint chat template, no thinking override. C/V: vLLM 0.17.1, Transformers 5.2.0.
Scoring	IFEval strict prompt-level all-instructions Boolean; GSM8K frozen tagged/last-number extraction and numeric normalization.	UGMathBench strict answer-slot extraction and type-aware judgment. The V extension loads the frozen scoring module.

Table 6: Recoverable model, sampling, and scoring settings for the two studies. Planned response counts and publisher-card training claims are distinguished from verified runtime records.

For Qwen, the checkpoint manifest binds the four exports by combined weight-hash prefixes `a6af60660f08`, `2ae5085b996c`, `5e7b7463dabc`, and `fce3793e4835` for t_0, t_1, t_2, t_3, respectively. Publisher cards describe a Qwen3-4B-Instruct-2507 student first distilled from 20,000 off-policy teacher responses, followed by 10,000 on-policy prompts and two increments of 5,000. The stated OpenThoughts-3 math slices are source positions 20k to 30k, 30k to 35k, and 35k to 40k. Accordingly, t_1, t_2, t_3 denote 10k, 15k, and 20k cumulative on-policy prompt exposures after t_0. They do not supply an optimizer-step axis.

The same cards describe a Qwen3-30B-A3B-Instruct-2507 teacher, reverse-KL distillation at temperature 1.0, and KDFlow/FSDP2 training for one epoch with batch size four and learning rate 2×10^{-6}. The stated rollout recipe uses SGLang, temperature and top-p of 1.0, four samples per prompt, an 8,000-token generation limit, and disabled thinking. These are training descriptions from the cards, distinct from the evaluation settings in Table 6. The archive lacks the exact trainer checkout and patches, immutable t_1 launcher, and optimizer, scheduler, RNG, and absolute-step save records. It therefore establishes the identities of the evaluated weights more strongly than continuous training lineage.

Study	Pool or cohort	Role and relation to other records
Llama	D	Historical discovery responses. Its hash-ordered $K = 8$ prefix selects 151 flips; larger nested prefixes change their labels.
	V_{128}	Separate responses for those fixed items; estimates directions without selecting the cohort.
	W_{512}	Another response pool for the same 151 items. Its depth subsets are nested within W, not within V.
Qwen	Greedy	Temperature-zero screen of 4,991 anchors; 171 fail at t_0 and succeed at t_1.
	C_{128}; C95	Separate confirmation responses; 95 of the 171 candidates pass the missing-robust direction rule. C95 names items, not a response pool.
	V	Follow-up responses withheld from confirmation. The post hoc reference extension uses 237 new and 19 historical endpoint cells.
	R128; R123	128 family-hash reference anchors. Five overlap with C95; removing them gives R123. Reference membership is not conditioned on confirmed improvement.
	S	Historical/common-environment response source for repeated measurements, covariate matching, and score sensitivity; distinct from C/V.
	Fresh	Archived 2,036-item response universe. It supplies broader summaries, not C95 confirmation or the later reference extension.

Table 7: Item cohorts and response pools used for selection, confirmation, and follow-up. The Llama and Qwen V pools are distinct; repeated measurements reuse items, not responses.

Scorer versions also remain separate. The Llama IFEval verifier is Google Research commit `08a8d6736475`, with result-independent fixed language-detector and Python RNG seeding in the repaired replay. GSM8K prefers the last tagged answer number, otherwise the last numeric expression, removes commas, and normalizes numeric text; absence of a candidate is a scored failure under that rule. The Qwen comparability ledger binds the UGMathBench judge source with SHA256 prefix `ac99a3048f3f` and its historical wrapper with `51e5b0af7e65`. The R128 extension explicitly loads the later frozen scoring module with prefix `249fb5fbcf05`. These identifiers refer to different software components and are not interchangeable scorer hashes. Exact hashes, prompt bindings, response IDs, and source versions are retained in the release manifests.

Holding out responses establishes their role relative to selection, not stochastic independence of every draw or invariance across runtime bundles. The historical request records bind prompt and decoding fields, but do not complete a matched cross-runtime response study. The separate common-environment S measurement is useful sensitivity evidence and does not close that gap. Accessible training-text checks likewise cover only the two later 5k prompt slices; they do not exclude exposure in unavailable earlier data or pretraining.

B Measurement Definitions and Historical Block Summaries

For an item with per-response success probability p under a fixed evaluation contract, K independent attempts produce at least one success with probability $1 - (1 - p)^K$. Learned and Forgotten, by contrast, are labels assigned to the observed before/after solved indicators. As the budget grows, both indicators can become one while the two endpoint probabilities remain unequal. Independence is useful for explaining this sampling relationship; it is not an established property of every saved generation stream.

Table 9 provides the archived $K = 8$ block view, while Table 8 lists all archived budgets. Their block partitions and Jaccard overlaps differ from the frozen hash-ordered D cohort followed in Section 3.2; the counts cannot be combined as if they came from the same partitions. In particular, the historical overlap does not establish a disjoint $K = 128$ confirmation or a per-item probability direction. The main directional comparison uses separate V_{128} responses for the fixed candidates and appears in Table 3. The formal archive also includes 2,000 MMLU-Pro items as a secondary measurement. Its archived response-level sample accuracy changes from 32.4051% for Native to 29.3191% for GAD.

Parser and scoring constraints make it a descriptive benchmark check here; it has no fixed-candidate independent-V direction analysis comparable to the IFEval and GSM8K main results.

Task	K	$\lvert L_A\rvert$	$\lvert L_B\rvert$	$\lvert F_A\rvert$	$\lvert F_B\rvert$	J_L	J_F
IFEval	1	42 (+0)	37 (+0)	99 (+0)	92 (+0)	0.145 (+0.000)	0.186 (+0.000)
	2	35 (-7)	35 (-2)	74 (-25)	80 (-12)	0.077 (-0.068)	0.213 (+0.027)
	4	25 (-17)	30 (-7)	68 (-31)	65 (-27)	0.122 (-0.023)	0.330 (+0.144)
	8	21 (-21)	25 (-12)	59 (-40)	54 (-38)	0.243 (+0.098)	0.449 (+0.263)
	16	20 (-22)	15 (-22)	43 (-56)	46 (-46)	0.250 (+0.105)	0.618 (+0.432)
	32	14 (-28)	9 (-28)	32 (-67)	39 (-53)	0.211 (+0.066)	0.578 (+0.392)
	64	13 (-29)	8 (-29)	24 (-75)	29 (-63)	0.312 (+0.167)	0.606 (+0.420)
GSM8K	1	141 (+0)	141 (+0)	165 (+0)	148 (+0)	0.119 (+0.000)	0.122 (+0.000)
	2	79 (-62)	98 (-43)	98 (-67)	103 (-45)	0.106 (-0.013)	0.086 (-0.036)
	4	55 (-86)	62 (-79)	61 (-104)	66 (-82)	0.136 (+0.017)	0.155 (+0.033)
	8	36 (-105)	38 (-103)	35 (-130)	46 (-102)	0.175 (+0.056)	0.141 (+0.019)
	16	29 (-112)	27 (-114)	18 (-147)	25 (-123)	0.167 (+0.048)	0.132 (+0.010)
	32	22 (-119)	20 (-121)	12 (-153)	16 (-132)	0.235 (+0.116)	0.167 (+0.045)
	64	11 (-130)	15 (-126)	6 (-159)	6 (-142)	0.300 (+0.181)	0.333 (+0.211)

Table 8: Archived Llama discovery results across K for IFEval and GSM8K. Red values are differences from the same task at $K = 1$; historical A/B blocks differ from the main $D/V/W$ cohort.

Task	Transition	Block A	Block B	Jaccard
IFEval	Learned	21	25	0.243
IFEval	Forgotten	59	54	0.449
GSM8K	Learned	36	38	0.175
GSM8K	Forgotten	35	46	0.141

Table 9: Observed transition counts and success-set Jaccard values at discovery budget $K = 8$ in the reported Llama blocks.

Task	Native (%)	GAD (%)	Difference (pp)
IFEval	62.97	53.01	−9.96
GSM8K	70.93	69.88	−1.05

Table 10: Full-population response accuracy for the archived Llama Native/GAD comparison. The main selected-item analysis uses separate V responses.

Task	Original D label	n	Supported	Opposite	Uncertain
IFEval	Learned	21	15	5	1
IFEval	Forgotten	54	48	2	4
GSM8K	Learned	41	20	8	13
GSM8K	Forgotten	35	13	6	16
Total		151	96	21	34

Table 11: Original, opposite, and uncertain W_{512} directions for the fixed 151 Llama $K = 8$ candidates. Each endpoint contributes 512 new responses; decisions use Beta$(1, 1)$ at 0.975.

The joint table in the main text is recomputed by a one-to-one join on task and item ID between the fixed-$K = 8$ cohort ledger and the corrected W_{512} direction ledger. All 151 IDs match, and every original discovery label agrees across the two files. No item is added, removed, or reselected. For the 21 opposite-direction W items, the signed empirical effect is $(c_{\text{after}} - c_{\text{before}})/512$ for originally

Learned items and its negative for originally Forgotten items. Its median is -5.66 percentage points; its range $[-17.77, -1.95]$ describes item heterogeneity, not uncertainty in a cohort mean.

Task	Fixed $K = 8$ candidates	Retained at $K = 64$	Retained and supported
IFEval	75	51	37
GSM8K	76	69	20

Table 12: The 151 D-selected Llama items followed from $K = 8$ to $K = 64$ without reselection. The last column counts Retained items with independent V_{128} support for the original direction.

C Qwen Coverage, Scoring, and Source Boundaries

The Qwen3-4B trajectory contains four fixed exports, t_0 through t_3. Historical greedy evaluation requested 14,973 UGMathBench items at each checkpoint. A separate Fresh response universe used a frozen 2,036-item subset with 128 responses per item and checkpoint. Tables 13 and 14 give the corresponding archived request and scored-response denominators. Neither Fresh nor greedy outcomes are relabeled as the separate C, V, or common-S/background measurements.

Historical generation used the checkpoint tokenizer's default chat template, the UGMathBench prompt builder, BF16 inference, and a 32,768-token output cap. Greedy decoding used temperature zero; Fresh used temperature 0.7 and top-$p = 0.95$. Probability estimates based on scored responses condition on binary correct/incorrect outcomes and keep nonbinary results separate. The main C/V decisions instead follow their own saved request manifests and the missing-score rules in Section 2. Reconstructing the direction-specific least-favourable completion agrees with all 513 published V labels; dropping unresolved scores changes 15 decisions. The source audit has itemwise counts and hashes, but no immutable receipt connecting the historical driver code to the execution output.

Checkpoint	K	Requested	Verified	Nonbinary
t_0	32	65,152	64,301	851
t_1	32	65,152	64,556	596
t_2	32	65,152	64,520	632
t_3	32	65,152	64,496	656
t_0	64	130,304	128,549	1,755
t_1	64	130,304	129,090	1,214
t_2	64	130,304	129,087	1,217
t_3	64	130,304	128,997	1,307
t_0	128	260,608	257,111	3,497
t_1	128	260,608	258,209	2,399
t_2	128	260,608	258,186	2,422
t_3	128	260,608	258,027	2,581

Table 13: Qwen Fresh scoring denominators for prefixes of 2,036 attempted items. Verified binary scores and nonbinary outcomes are shown separately.

Checkpoint	Requested	Verified	Success (%)
t_0	14,973	14,342	55.62
t_1	14,973	14,706	55.66
t_2	14,973	14,745	55.21
t_3	14,973	14,718	55.63

Table 14: Qwen greedy success rates among responses with verified binary scores. Nonbinary outcomes are excluded from the displayed denominators.

Fixed cohort	Checkpoint	Supported	Opposite	Uncertain	Total
All screened candidates	t_1	106	9	56	171
All screened candidates	t_3	82	24	65	171
C-confirmed subset	t_1	86	0	9	95
C-confirmed subset	t_3	60	9	26	95

Table 15: Independent Qwen V direction decisions relative to t_0. Supported, uncertain, and opposite counts refer to baseline comparisons at each later checkpoint.

C.1 Execution and output-limit checks

Qwen checkpoint responses came from multiple execution owners. Request ownership was checked without canonical conflicts, but a cross-owner response replay was not completed. A separate precision comparison also changed the attention backend, and the temperature control had no frozen scientific pass/fail threshold. These checks qualify the execution record; they do not add training replications.

The archived output-limit control selected 1,533 length-capped requests without using correctness and raised the response allowance from 32,768 to 40,960 tokens. Of those requests, 1,528 terminated normally and five remained technically missing. Enumerating all 32 binary completions of those five scores bounded improvement within the selected requests. The control addresses response length and missingness, not the cause of a longitudinal change. Table 16 reports its request counts.

Checkpoint	Fresh	Greedy	Total	Fresh missing
t_0	268	462	730	0
t_1	256	107	363	4
t_2	118	75	193	0
t_3	165	82	247	1
Total	807	726	1,533	5

Table 16: R3 selected requests by checkpoint and branch. All five recorded technical missing scores occur in Fresh.

D Family Populations and Time Estimands

The related-item comparison is a supporting analysis. The family-equal fixed-denominator sibling-control change is $+5.11$ percentage points in the 95 families whose anchor has a confirmed improvement. Native171 and the strict operational subset have positive points as well, but only the confirmed group has a positive outer sensitivity envelope. These recorded family relations do not establish transfer to new problems.

Fixed cohort	Fixed-n change	Unknown bounds
All 171	-7.85	$[-8.50, -7.23]$
C95	-12.25	$[-12.43, -12.07]$

Table 17: Qwen fixed-denominator V changes across checkpoints, in percentage points. Bounds assign identified unknown scores their most favorable or adverse outcomes.

Family	Fixed-n change	Outer envelope
Native$_{171}$	$+3.37$	$[-1.43, +8.02]$
C-confirmed$_{95}$	$+5.11$	$[+1.45, +8.63]$
Strict dual-AI$_{84}$	$+3.29$	$[-3.28, +9.81]$

Table 18: Qwen family-level changes from t_0 to t_3 in percentage points. Points use fixed denominators; envelopes include missing-score and family-dependence sensitivity.

For each family f, let (s_{fj}, c_{fj}), $j = 1, 2$, denote its two recorded sibling-control pairs. The conditional point estimate averages

$$D_f = \frac{1}{2}\sum_{j=1}^{2}\left[(\mu_{s_{fj},t_3} - \mu_{s_{fj},t_0}) - (\mu_{c_{fj},t_3} - \mu_{c_{fj},t_0})\right] \tag{3}$$

across families, where $\mu = (c+1)/(c+w+2)$. Its values are $+3.11$, $+4.97$, and $+2.71$ percentage points for the three populations below. For the main table, we instead replace every μ by $c/128$, so the point and missingness sensitivity use the same score definition. Table 18 reports three fixed populations: all 171 screened candidate families, the 95 whose anchor passed C confirmation, and 84 accepted by the recorded strict AI structural screen. The last criterion is an operational screen, not human validation of mathematical equivalence.

The accompanying sensitivity envelope uses the same fixed-denominator contrasts as the revised point estimates. These points keep only frozen correct counts in the numerator; the envelope additionally completes unknown outcomes. For each of 10,000 replicates, the archived dependency graph's connected components are sampled with replacement. All families within a selected component retain their multiplicity, and the resampled family mean is normalized by the resulting family count. Coefficients for each shared item-checkpoint score are combined before unknown outcomes are set to zero or one, yielding lower and upper values L_b, U_b. The envelope is $[Q_{0.025}(L_b), Q_{0.975}(U_b)]$, with deterministic seeds derived from master seed 20260922 and the population identifier. It combines missing-outcome completion with the chosen component resampling scheme. No calibrated simultaneous coverage over an underlying family population has been established, and the envelope combines completion and resampling uncertainty instead of being a response-sampling interval for these fixed families.

The archived Qwen family analysis used a frozen match and three matching variants across 171 families. Table 19 contains its descriptive means. That analysis has an earlier population and estimand definition than the three populations and conditional Beta/completed-data analyses in Appendix D. Its quantiles summarize variation between families, not uncertainty around the mean. Neither analysis supplies independent human validation of structural equivalence, and the 95 confirmed families cannot stand in for a benchmark-wide family population.

Matching	Pairs	Sibling change	Control change	Difference
Main	342	5.14	1.66	3.48
Nearest	342	5.14	0.72	4.42
Strict	341	5.12	1.58	3.53

Table 19: Family-weighted Qwen sibling and control changes from t_0 to t_3 under three matching rules. Values are percentage points; each rule includes 171 families.

The post hoc V time analysis directly pairs Qwen items at t_1 and t_3. Item-equal conditional Beta differences are -7.789 percentage points for all 171 and -12.063 for the confirmed 95; these item sets are fixed rather than independent training runs. The frozen fixed-denominator changes are -7.849 and -12.245 points, respectively (Table 17). Their accompanying bounds place unresolved outcomes at their allowable extremes. The archived time-results file also contains a bounds-midpoint field; we recompute the displayed correct yield from the itemwise numerators and do not substitute that midpoint for the frozen score. Because historical V responses span runtime bundles, the common-environment S result remains a separate sensitivity measurement.

D.1 Empirical response-budget curves

Tables 20 and 21 give every saved budget from $K = 1$ to 128 for the historical S and separate V responses. These values draw without replacement from the same 128 responses per item and checkpoint. In V, C95 falls from 79.24% to 67.00% at $K = 1$, yet both endpoints reach 100% empirical pass@128 because every selected item has at least one scored success in each block. R123 behaves differently: its t_3 empirical pass@128 is 67.48%. Thus any-success coverage and per-response accuracy answer different questions on these fixed items; the S and V columns are separate response measurements, not training replications.

K	C95 t_1	C95 t_3	R128 t_1	R128 t_3	R123 t_1	R123 t_3
1	78.63	67.07	57.90	57.41	56.43	55.92
2	90.58	80.23	60.62	59.79	59.05	58.19
4	96.39	89.18	62.26	61.22	60.72	59.64
8	98.85	94.49	63.85	62.59	62.38	61.07
16	99.76	97.31	65.74	64.22	64.34	62.77
32	99.98	98.68	67.73	66.11	66.42	64.73
64	100.00	99.41	69.45	68.32	68.21	67.03
128	100.00	100.00	71.09	71.09	69.92	69.92

Table 20: Empirical pass@K in saved Qwen S responses, averaged over items. Each value draws K from 128 responses without replacement; R123 excludes five C95-overlap items.

K	C95 t_1	C95 t_3	R123 t_1	R123 t_3
1	79.24	67.00	56.19	55.89
2	90.99	80.08	58.93	58.28
4	96.58	89.13	60.65	59.84
8	98.97	94.58	62.34	61.23
16	99.82	97.46	64.22	62.74
32	99.99	98.86	65.99	64.41
64	100.00	99.46	67.29	66.14
128	100.00	100.00	68.29	67.48

Table 21: Empirical pass@K in saved Qwen V responses, averaged over items. Each value draws K from 128 responses without replacement; R123 excludes five C95-overlap items.

E Response Blocks in Separate Settings

Setting 2 is a Qwen 1.7B Base/Post OPD comparison, distinct from the four-checkpoint Qwen3-4B trajectory. D2 generated and scored three new full-universe $K = 256$ blocks: 130 items, two endpoints, three blocks, and 256 responses per block give 199,680 scored responses. Table 22 records four-way labels separately for each block. The pooled directional analysis instead uses all three blocks and a single Holm correction over 260 one-sided Fisher hypotheses, yielding 89 positive, three negative, and 38 uncertain items. Old B_0 responses are excluded from that primary test. Of 22 earlier selected transitions, eight retain their original exact label across all three new blocks. The 81 earlier signed direction declarations all have the same D2 direction; this comparison is with old declarations, not a new training run.

Block	Retained	Both unsolved	Forgotten	Learned
B_1	86	20	7	17
B_2	83	21	5	21
B_3	83	21	5	21

Table 22: Observed four-way labels for all 130 D2 items in three new Base/Post blocks at $K = 256$. Blocks are separate response samples, not directional tests.

The separate Llama-3.2 GRPO/RLVR MATH-500 study uses 436 formal items. Its B_0 $K = 64$ labels comprise 362 Retained, 42 Both unsolved, 16 Forgotten, and 16 Learned items. Three fresh $K = 64$ blocks then follow the 32 selected transitions. Only one of the 32 retains its original label in all blocks, while 15 exhibit an opposite label at least once. Using only fresh responses, a Jeffreys-Beta posterior at the specified 0.95 threshold finds three original-direction, two opposite-direction, and 27 uncertain items. The selected-cohort result does not estimate transition prevalence over all 436 items. Batch-level random-number generation and selection into the 32 limit a direct numerical comparison with Setting 2 or the main Llama Native/GAD study.

Reanalysis under the main statistical rules. To isolate the contribution of the different inferential rules, we recompute the separate-study counts with the main Beta$(1, 1)$ prior and 0.975 threshold. MATH-500 pools only B_1, B_2, B_3, giving 192 binary responses per endpoint for each of its 32

Setting	Label n	Same	Changed	Direction n	Positive/original	Negative/opposite	Uncertain
Qwen 1.7B D2	22	8	14	130	89	3	38
Llama MATH-500	32	1	31	32	3	2	27

Table 23: Label recurrence and independent direction checks in separate Qwen and Llama settings. Same counts labels persisting across three fresh blocks; direction columns use each setting's stated rule.

candidates. One original direction is supported and 31 are uncertain; none is opposite. Fisher/Holm over all 64 one-sided hypotheses leaves all 32 uncertain. The historical Jeffreys-0.95 counts above are retained as the original analysis, rather than being silently substituted for the stricter result.

For D2, we retain the full 130-item family and its 260 hypotheses, then inspect the 22 flips selected before those responses were generated. The common Beta rule supports 12 original directions and leaves ten uncertain; the full-family Fisher/Holm result supports 11 and leaves 11 uncertain. Neither rule supports an opposite direction (Table 24). These outcomes show that the corrected reversals in the primary Llama cohort are not reproduced in every available setting. They do not establish equal underlying reliability across settings, whose discovery budgets, tasks, and generation procedures differ.

Selected cohort	n	Depth	Rule	Supported	Opposite	Uncertain
Llama MATH-500	32	192	Beta$(1,1)$	1	0	31
Llama MATH-500	32	192	Fisher/Holm	0	0	32
Qwen 1.7B D2	22	768	Beta$(1,1)$	12	0	10
Qwen 1.7B D2	22	768	Fisher/Holm	11	0	11

Table 24: MATH-500 and D2 selected cohorts reanalyzed with Beta$(1,1)$ at 0.975 and Fisher/Holm at 0.05. Holm covers both directions of all 32 MATH-500 or 130 D2 candidates.

An earlier Qwen OPD Setting 1 report lists 17 selected transitions at $K = 1024$, three of which kept their label across three fresh checks. Its per-item scores and selected manifest were not restored in this review package. We retain that figure as historical summary-level context and do not combine it with the complete D2 or MATH-500 itemwise analyses.

F Known-Truth and Same-Model Reporting Checks

The same-model controls in Table 25 compare nonoverlapping response blocks at matched budgets. Each pair holds the model weights fixed, so a transition arises from finite sampling under the saved inference and scoring rules. This is a concrete null reference for the label, whereas the full-budget Fisher/Holm analysis addresses a different directional declaration. The controls do not adjudicate the cause of the Base/Post differences in either separate setting.

Setting	Same-model pair	Items	M0 transitions	M1 directions
Qwen Setting 2, $K = 32$	Base A/B	130	14	0
Qwen Setting 2, $K = 32$	Post A/B	130	8	0
Llama MATH-500, $K = 8$	Base A/B	436	58	0
Llama MATH-500, $K = 8$	Post A/B	436	61	0

Table 25: Same-model controls from nonoverlapping response blocks at matched budgets. M0 counts success-set transitions; M1 counts Fisher/Holm directions at full budget.

The reporting-rule audit fixes five rules on development simulations before evaluating them on separate validation universes. At each item and endpoint it generates independent binomial discovery counts with $B/4$ trials and confirmation counts with $3B/4$ trials; their sum supplies the full-budget rules. M0 declares the direction of a full-budget binary flip. M1 applies one-sided Fisher tests with

Holm correction over both directions of all N items. M2 thresholds a Jeffreys-prior direction probability at 0.95. M3 selects discovery flips and applies Fisher/Holm to both directions of only the selected items using confirmation counts; M4 applies the Jeffreys rule to those confirmation counts. Thus M3 and M4 reserve responses for confirmation, but use different inferential rules from the main empirical Beta$(1, 1)$ analysis.

In S0, equal endpoint probabilities cycle through 0, 0.0001, 0.001, 0.01, 0.05, 0.1, 0.3, 0.7, 0.95, and 1. S4 cycles baseline probabilities through 0.0001, 0.001, 0.01, 0.05, 0.1, 0.3, 0.7, and 0.95, assigns 60% unchanged items and 20% to each sign, cycles nonzero magnitudes through 0.005, 0.01, and 0.05, and clips changed probabilities to $[0, 1]$. S5 cycles endpoint pairs $(0.1, 0.3)$, $(0.3, 0.1)$, $(0.05, 0.2)$, and $(0.7, 0.5)$. The item false-positive rate divides false declarations by unchanged items; family error records whether any false or wrong-sign declaration occurs in a simulated universe; correct-direction rate divides correct signed declarations by changed items. These definitions explain the different denominators below. Table 26 now reports all five rules in the matched $N = 130$, $B = 256$ validation slice, averaged over 2,000 universes, for the no-change S0, mixed S4, and detectable-change S5 scenarios. Its item false-positive rate, family error frequency, and correct-direction rate have different denominators. The archived M2 and M4 decisions approximate Jeffreys-Beta direction probabilities by matching moments to a normal distribution. The narrower numerical audit in Table 27 replaces only decisions within a specified threshold band by exact integration, so neither table establishes a fully exact-Beta evaluation across every scenario. The rates belong to the simulation's data-generating model, not the empirical OPD populations.

Frozen rule	S0 false positive	S0 family error	S4 correct direction	S4 false positive	S5 correct direction
M0: binary solved label	0.05390	1.0000	0.17424	0.06955	< 0.00001
M1: full-count Fisher/Holm	0.00014	0.0185	0.04848	0.00011	0.95797
M2: full-count Jeffreys	0.05478	0.9995	0.23676	0.06863	0.99968
M3: split exact confirmation	0.00002	0.0030	0.03245	0.00003	0.00971
M4: split Jeffreys confirmation	0.00291	0.3175	0.07103	0.00371	0.00977

Table 26: Synthetic calibration of five archived rules with 130 items, 256 responses per endpoint, and 2,000 validation universes. S0 has no changes, S4 mixes changed and unchanged items, and S5 has detectable changes; M2 and M4 use the archived normal approximation.

Rule	S0 false pos.	S0 family error	S4 correct dir.	S4 false pos.	S4 family error	S5 correct dir.
M2	0.06287	1.0000	0.26225	0.07969	0.9995	0.99968
M4	0.00638	0.5665	0.08102	0.00829	0.4925	0.00977

Table 27: Partial exact-Beta correction of M2 and M4 in the archived simulations. Exact integration replaces the normal approximation only within 0.10 posterior probability of each threshold.

The archived stage ledger decomposes M3's pooled rate directly (Table 28). In S5, only 2,543 of 260,000 changed-item observations enter the candidate set; 2,524 are correctly confirmed. Selection probability is therefore 0.978%, confirmation conditional on selection is 99.253%, and their product is the reported 0.971%. Its low overall rate is predominantly an admission bottleneck. S4 behaves differently: 15.549% enter and only 20.871% of entrants are correctly confirmed. Both stages limit its power. The M3 S5 correct-direction rate of 0.00971 pools all changed items in that validation

Scenario	Changed	Selected	Correct	Admission	Correct | selected	Overall
S4	104,000	16,171	3,375	15.549%	20.871%	3.245%
S5	260,000	2,543	2,524	0.978%	99.253%	0.971%

Table 28: M3 candidate admission and conditional direction confirmation in the S4 and S5 simulations. Counts pool 2,000 universes with 130 items and 256 responses per endpoint.

scenario. Within the narrower $|\Delta p| = 0.15$ and $0.01 \leq p < 0.10$ stratum, M3 makes 2,371 correct declarations among 64,000 item observations, a rate of 0.03705. This conditional slice has a different denominator, and no matched M1 stratum ledger is available for a direct numerical comparison.

G Historical Reference Measurement and Score Sensitivity

The completed Qwen common-S/later-background analysis includes fixed C95 and R128 groups at t_1 and t_3. Under the original fixed-plan machine score, their mean changes are -11.554 and -0.488 percentage points, respectively; the shared-item-aware direct difference is -11.066 points (Table 29). The groups overlap on five items. These retrospective measurements use a response source distinct from the later completed V comparison in Section 3.4. Neither source identifies why C95 was selected or why its later responses changed.

Fixed group or contrast	Items	$t_3 - t_1$ change
C95	95	-11.554
R128	128	-0.488
C95 minus R128	N/A	-11.066

Table 29: Fixed-plan t_1 to t_3 machine-score changes in the common S and later background responses. C95 and R128 share five items; values are percentage points.

For the historical S C95 responses, the item-equal conditional Beta means at t_0, t_1, t_3 are 54.50%, 78.37%, and 67.00%, whereas the corresponding frozen fixed-denominator correct yields are 54.13%, 78.63%, and 67.07%. The fixed-denominator S reference yields are 57.90% to 57.41% for all R128 and 56.43% to 55.92% for the nonoverlapping R123. These historical S numbers are kept separate from the later independent V extension and from its conditional Beta means. For this historical S measurement, setting each already marked scoring-path or unresolved response to either binary outcome once gives a deterministic C95 change range of $[-12.20, -11.04]$ percentage points and a shared-item-aware C95 minus R128 range of $[-13.55, -9.04]$ points. These path-union bounds concern marked records, use the historical S responses, and cannot be applied to undiscovered semantic errors in V.

The V extension's fixed planned counts are shown in Table 30. All C95 and R128 items have 128 responses per endpoint. One identified scorer execution failure was recovered as incorrect after the original score freeze, with no change to the correct numerators or the reported group contrast. The deterministic missing-outcome bounds in Section 3.4 concern only the remaining identified unresolved scores.

Group	Checkpoint	Planned	Correct	Incorrect	Unresolved
C95	t_1	12,160	9,636	2,502	22
C95	t_3	12,160	8,147	3,992	21
R128	t_1	16,384	9,434	6,806	144
R128	t_3	16,384	9,391	6,789	204
R123	t_1	15,744	8,846	6,754	144
R123	t_3	15,744	8,799	6,742	203

Table 30: Audited fixed-denominator V counts for C95, R128, and nonoverlapping R123. One R123 t_3 unknown was confirmed incorrect without changing correct counts.

Three supplementary checks probe what the group difference can and cannot mean. First, a no-change simulation holds the fixed item membership and the scorer constant while setting each item's t_1 and t_3 response probability equal to its pooled V estimate (Table 31). The C95 decline lies beyond this response-sampling null in 20,000 replicates, whereas the R123 change gives a one-sided Monte Carlo $p = 0.05495$. This is not a no-training control. Second, direct S versus V repeats at the same checkpoint show small C95 mean differences (Table 32); the R128 overlap subsets are too small and differ across checkpoints. Third, Table 33 shows the historical full, mutually exclusive, and common-support reference comparisons under their distinct S-source budgets and weighting rules. The restricted common-support group and the post hoc selection proxy do not recreate C95's greedy-to-C selection rule for R128, so these checks remain descriptive.

Fixed group	Items	Observed change (pp)	Null mean (pp)	One-sided p
C95	95	-12.25	-0.004	0.00005
R123	123	-0.30	-0.002	0.05495

Table 31: Fixed-item no-change simulation for the Qwen V comparison. Endpoints share each item pooled response probability across 20,000 Bernoulli replicates.

Group	Checkpoint	Shared items	S (%)	V (%)	$V - S$ (pp)
C95	t_0	95	54.13	54.43	+0.30
C95	t_1	95	78.63	79.24	+0.62
C95	t_3	95	67.07	67.00	-0.07
R128	t_0	11	79.69	80.68	+0.99
R128	t_1	8	89.75	90.14	+0.39
R128	t_3	11	94.60	92.90	-1.70

Table 32: Fixed-denominator Qwen scores in historical S and separate V responses at each checkpoint. Each row uses items shared by both pools; the R128 subset varies by checkpoint.

Comparison	C95 n	Ref. n	C95 Δ	Ref. Δ	Difference
Full S, 128	95	128	-11.55	-0.49	-11.07
Evaluation S, 64	95	128	-11.25	-0.45	-10.80
Exclusive, 64	95	123	-11.25	-0.51	-10.74
Common support, 64	47	77	-11.47	-0.49	-10.98
Weighted support, 64	47	77	-12.13	-0.61	-11.52

Table 33: Retrospective S-source C95 and reference changes in percentage points. Rows use full or evaluation budgets, then remove overlap and restrict or reweight common support.

The post hoc R128 DESIGN-threshold proxy retains 114 to six items as its threshold grows (Table 34). Its selection rule lacks the greedy discovery and independent C confirmation used for C95, so these changes cannot be interpreted as a selection-matched control.

Post hoc DESIGN threshold	Items	DESIGN later change (pp)	EVALUATION later change (pp)
$t_1 - t_0 \geq 0.00$	114	-1.14	-0.62
$t_1 - t_0 \geq 0.05$	21	-5.28	-4.24
$t_1 - t_0 \geq 0.10$	12	-6.25	-4.82
$t_1 - t_0 \geq 0.15$	6	-5.99	-4.43
$t_1 - t_0 \geq 0.20$	6	-5.99	-4.43

Table 34: Post hoc R128 selection by early DESIGN response change and later t_3 to t_1 change in DESIGN and EVALUATION responses. This proxy does not repeat the greedy plus C confirmation used for C95.

Replay of the historical improvement selector. The original anchor, early-acquisition, and reference manifests identify all 128 reference items within the same 4,991-anchor frame. Exactly eight belong to the 171 greedy candidates. Joining the archived C decisions gives five confirmed reference items, two uncertain items, and one opposite item. All five confirmed reference items already belong to C95. Thus replaying the same historical selector yields no nonoverlapping confirmed reference cohort after the shared items are removed. The release includes all 128 membership rows and the input hashes. This is a replay of recorded decisions, not a new confirmation sample; a separate comparison requires a distinct sampling or experimental design.

A shared-entry comparison within the 171 candidates. An additional comparison holds the initial greedy-entry criterion fixed. All 171 anchors passed that screen before the separate C mea-

surement; C95 then passed confirmation and the remaining U76 did not. We join their recorded C decisions to the untouched V endpoint counts. U76 decreases from 61.19% to 58.84%, a -2.35-point change, compared with -12.25 for C95. Their difference is -9.89 points. Under the fixed-item correct-yield posterior, 100,000 independent draws with seed 2026092603 give a contrast mean of -9.74 and equal-tailed 95% interval $[-11.14, -8.33]$. Completing identified unknowns instead gives deterministic bounds $[-11.24, -8.48]$ for the empirical contrast.

Table 35 reports every original confirmation branch. Both confirmed branches decline, by 13.30 points for the 79 complete-score confirmations and 7.03 for the 16 missing-robust confirmations. The ten candidates assigned an opposite C direction instead increase by 8.98 points in the later V comparison. This association across separately measured stages is informative about the selected items, but the confirmation branches differ in measured early change and endpoint levels. U76 is therefore a shared-entry comparison, not a same-confirmation reference or evidence that confirmation itself causes later decline.

Candidate group	n	$V : t_1$	$V : t_3$	Change	Unknown bounds
All candidates	171	71.22	63.37	-7.85	$[-8.50, -7.23]$
Confirmed (C95)	95	79.24	67.00	-12.25	$[-12.43, -12.07]$
Not confirmed (U76)	76	61.19	58.84	-2.35	$[-3.59, -1.18]$
Complete-score confirmation	79	81.02	67.72	-13.30	$[-13.31, -13.27]$
Missing-robust confirmation	16	70.46	63.43	-7.03	$[-8.06, -6.15]$
Complete-score uncertainty	44	68.71	64.70	-4.01	$[-4.03, -3.96]$
Complete-score opposite	10	56.02	65.00	$+8.98$	$[+8.98, +8.98]$
Missing-sensitive / uncertain	22	48.51	44.32	-4.19	$[-8.42, -0.25]$

Table 35: Fixed-denominator V scores for the 171 greedy-selected Qwen candidates and their C confirmation strata. Endpoint scores are percentages, changes are percentage points, and bounds complete identified unknown scores.

Post hoc adjustment with separate response covariates. We use the historical S baseline correct yield and S early change, each based on 128 responses, together with exact task-domain agreement. A one-to-one match without replacement first maximizes the number of pairs, then minimizes the sum of absolute differences in those two covariates. The allowable difference on each covariate is five, ten, or 20 percentage points. All three choices were recorded before this computation and all are reported; they were selected after the original experiment. The separate V endpoint changes are not used in matching. C95 coverage remains low, and the contrasts change materially with the allowed imbalance (Table 36). These restricted contrasts do not select reference membership using the greedy-to-C rule and cannot identify a selection effect.

Caliper	Pairs	C95 coverage	S baseline gap	S gain gap	C95 V change	Contrast
5 pp	5	5.26%	$+0.47$	$+1.72$	-12.66	-9.38
10 pp	16	16.84%	-2.29	$+3.22$	-3.86	-2.73
20 pp	35	36.84%	-6.56	$+8.50$	-7.61	-5.74

Table 36: Post hoc matching of nonoverlapping C95 and R123 items on S baseline, early gain, and domain. Gaps and V change contrasts subtract the matched R123 mean from C95, in percentage points.

An earlier scoring audit reviewed 384 Qwen full answers through two frozen AI channels, including 256 probability-sampled answers and 128 targeted diagnostic answers. Both channels called 45 originally machine-incorrect responses mathematically correct in the probability sample, but some involved format or gold disputes and the channels did not constitute independent human adjudication. An expanded A review then locked judgments for 1,799 distinct answers, and the corrected frame replaced nine cases while preserving the sample size and A label totals. The nine added cases were judged after the evaluator knew their machine-correct stratum, so they are recorded as an addendum rather than strictly blind first-pass judgments. The B review agrees with A on the corrected sample. Among machine-incorrect responses in A's original review, 352 were called semantically

correct, yet semantic agreement does not itself determine whether the frozen answer-slot, format, tolerance, and gold rules should accept them. No executable response-wide correction rule or human ground truth follows from the two AI reviews.

We use the four sampled machine-incorrect V strata to illustrate one limited sensitivity calculation. Each stratum contains 128 reviewed responses. Weighting its A semantic-correct fraction by the corrected stratum population and hypothetically accepting those responses gives a C95 minus R123 change difference of -12.23 percentage points; excluding sampled cases flagged for format or gold gives -14.03 points (Table 37). Both values remain negative, but this exercise holds machine-correct and technically unknown outcomes fixed, assumes the sampled AI labels extend within each stratum, and is not a formal rescoring or a deterministic bound on undiscovered semantic errors. The frozen score and C95 membership remain unchanged, and the main contrast should not be described as human-validated mathematical correctness.

Scoring scenario	C95	R123	C95 minus R123
Frozen machine score	-12.25	-0.30	-11.95
Accept all A semantic-correct	-9.23	$+3.00$	-12.23
Accept only without format or gold flags	-10.70	$+3.34$	-14.03

Table 37: Hypothetical Qwen V changes after applying A review fractions to machine-incorrect strata. Machine-correct and unknown scores stay frozen; entries are percentage-point changes, not adjudicated scores.

A targeted human-review preparation ranks C95 items by their contribution to the frozen V decline. The ten largest contributors account for 717 of the net 1,489 lost scored successes, or 48.15%. The supplied semantic sample contains 52 of their 2,560 endpoint responses. We prepare those available answers with model, checkpoint, and prior-label metadata removed, and retain a separate identity key. This subset does not cover the ten items fully and has not received human adjudication. The 21 Llama reversals are also enumerated, but the available W delivery supplies counts without the corresponding full-response review packet. A complete assessment of capability, formatting, extraction, and gold disputes remains outstanding.

G.1 Recovered original responses and execution-worker checks

The nested common-S source archive contains 48,640 original responses for all 95 C95 items at all four checkpoints. We verify 11,429 file hashes and 5,694 raw-artifact receipts, then join every response to its frozen score and planned request. Raw-text hashes, canonical record hashes, checkpoint identity, item, sample index, seed, and prompt hash agree. The later endpoints contain 24,320 responses; their send-intent records additionally agree with the recorded execution worker and GPU. This recovery concerns S, not the primary V or Llama W responses, and verifies data linkage rather than mathematical truth.

The rebuilt S counts are 9,561 correct at t_1 and 8,156 at t_3, each out of 12,160, giving a -11.554-point change. The 29 and 27 unknown outcomes imply bounds $[-11.793, -11.332]$. Sixty-five items decline in S, and 63 decline in both S and V. Across all 95 paired item changes, Spearman correlation is 0.947. The ten largest contributors to the V decline all decline in S as well, with a mean S change of -52.34 points. Their full 2,560 S endpoint answers are now supplied in a separate packet with model, checkpoint, and prior-score fields removed. That packet expands access to a second response source; it does not increase the 52 available primary-V answers or represent completed human adjudication.

Actual worker assignments permit a check beyond arbitrary response partitions. All eight workers contain both endpoints for every C95 item. Within each worker we calculate each item's $t_3 - t_1$ rate difference using that worker's endpoint denominators, then average equally across the 95 items. Every worker has a negative mean, ranging from -12.80 to -8.95 points. Repeating the calculation after omitting each worker gives -11.96 to -11.42 (Table 38). These overlapping checks show that the S decline is not confined to one execution worker. They neither establish iid responses within a worker nor calibrate uncertainty for the primary V/W pools.

Worker	Endpoint responses	Worker-only change	Omit-worker change
0	2954	−12.06	−11.53
1	3126	−8.95	−11.96
2	2882	−12.62	−11.46
3	2900	−12.80	−11.42
4	3168	−11.53	−11.61
5	3106	−12.04	−11.50
6	3024	−11.21	−11.52
7	3160	−12.39	−11.44

Table 38: C95 changes in recovered common-S responses by recorded worker. Each row spans all 95 items; rates average itemwise endpoint fractions and changes are percentage points.

The recovered executor submits asynchronous requests and publishes completed records in chunks of 16 or after a time threshold. A raw-file shard is consequently a storage unit, not a physical generation batch, and we do not treat shard membership as such. The frozen scoring-path counts provide a different diagnostic (Table 39). Official-strict correct counts fall by 1,420 while fallback correct counts increase by 15, summing to the 1,405 net lost successes. Thus disappearance of fallback-accepted responses does not account for this S decline. The strict path can still contain semantic, gold, or extraction disagreements; no response is reclassified by this decomposition.

Extraction path	t_1 records	t_1 correct	t_3 records	t_3 correct	Yield change
Official strict	12098	9543	12084	8123	−11.678
Balanced-box fallback	61	18	76	33	+0.123
No parser result	1	0	0	0	+0.000
All paths	12,160	9,561	12,160	8,156	−11.554

Table 39: Frozen scoring paths in recovered S responses. Contributions are path-specific changes in correct counts divided by the full 12,160-response endpoint denominator.

G.2 Concentration of the selected-group decline

The 95 C95 items include 68 negative, six zero, and 21 positive observed $t_3 - t_1$ changes. Their median change is -7.03 percentage points. R123 instead has 23 declining, 78 tied, and 22 increasing items, with median zero. These distributions further distinguish the fixed groups without explaining their different selection histories.

To quantify the influence of a few large declines, we rank C95 by its lost correct counts, break ties by item ID, and remove the first one, five, ten, or 20 items. All four choices are retrospective. The ten largest contributors account for 717 of 1,489 net lost successes, yet the remaining 85 items lose 772 successes across 10,880 responses per endpoint. Their mean change is -7.10 points, with unknown-score bounds $[-7.30, -6.90]$. After removing 20 items, the remaining mean is -3.86 points. These restricted estimands describe concentration and do not estimate an unbiased population effect after trimming.

Removed	Remaining n	Mean (pp)	Median (pp)	Down/tie/up	Unknown bounds (pp)
0	95	-12.25	-7.03	68/6/21	[-12.43,-12.07]
1	94	-11.44	-6.64	67/6/21	[-11.62,-11.26]
5	90	-9.15	-6.25	63/6/21	[-9.34,-8.97]
10	85	-7.10	-4.69	58/6/21	[-7.30,-6.90]
20	75	-3.86	-3.12	48/6/21	[-4.06,-3.65]

Table 40: C95 V changes after removing the largest observed decline contributors. Removal sets are chosen from the data; means, medians, and unknown-score bounds describe the remaining items.

Twelve of the 14 recorded C95 domains have negative mean changes. The two positive rows each contain one item, and several other strata are also small. Removing each domain in turn gives C95 changes from -12.71 to -11.35 points (Table 41). Thus neither the ten largest contributors nor any single recorded domain accounts for the full group decline. These checks cannot determine whether frozen-score changes reflect reasoning, formatting, extraction, or gold disputes.

Domain	C95 n	C95 change	R123 n	R123 change	C95 without domain
Algebra	13	-17.91	9	-0.26	-11.35
Arithmetic	2	-1.56	8	-0.10	-12.47
Multivariable calculus	12	-11.78	17	+1.29	-12.31
Single-variable calculus	28	-11.13	23	-2.31	-12.71
Complex analysis	1	+8.59	1	+0.00	-12.47
Differential equations	3	-2.34	5	+0.00	-12.57
Financial mathematics	5	-16.72	8	+3.81	-12.00
Geometry	1	-0.78	2	+0.00	-12.37
Linear algebra	11	-9.87	12	-1.11	-12.56
Number theory	1	-33.59	1	+0.00	-12.02
Probability	8	-16.11	8	+0.20	-11.89
Set theory and logic	1	+5.47	2	-0.78	-12.43
Statistics	6	-7.81	11	+2.06	-12.54
Trigonometry	3	-26.30	10	+0.08	-11.79

Table 41: C95 and R123 V changes by recorded domain, in percentage points. The last column recomputes the C95 mean after removing that domain.

G.3 Sensitivity to adverse binary-score changes

For each corrected Llama reversal, we search integer allocations of score changes between its two endpoints, moving only against the observed difference. All other items remain fixed, and Fisher/Holm is recomputed over all 302 hypotheses for every allocation. Table 42 gives the smallest number that removes that item's adjusted support and separately the number needed to erase its observed sign. Each minimum has an explicit count-level witness, and every allocation with one fewer change is checked. The former minima range from five to 46 of 1,024 response scores; the latter range from 31 to 91. These hypothetical label changes are neither detected scoring errors nor semantic judgments. They identify the amount and concentration of correction to which each conclusion is sensitive.

Item	Lose Holm support	Fraction of responses	Erase observed sign
IFEval 2785	5	0.49%	31
IFEval 2941	46	4.49%	91
IFEval 3327	9	0.88%	41
IFEval 3362	5	0.49%	41
IFEval 3456	7	0.68%	32
GSM8K 1094	11	1.07%	53
GSM8K 1129	19	1.86%	66
GSM8K 1145	12	1.17%	47
GSM8K 1174	26	2.54%	63

Table 42: Minimum adverse binary-score changes needed to remove Holm support or erase the observed sign for each of nine W reversals. Holm is rerun over all 302 hypotheses.

For the frozen C95 mean, 1,489 net favorable binary-label changes erase the aggregate decline. The C95 minus R123 contrast reaches zero after at least 1,453 optimally placed changes: one C95 score contributes $1/12{,}160$ to the contrast, more than the $1/15{,}744$ contribution from one R123 score. Changing 1,453 C95 t_3 failures to successes is a feasible count-level witness. Membership is held fixed, and the identified unknown scores are not relabeled in this construction. These minima do not estimate error prevalence or make any claim about which corrections are mathematically justified. Targeted human adjudication remains necessary for that interpretation.

H Additional Training and Structural Diagnostics

FitFunc is a completed 160-step on-policy run with a Qwen3-1.7B student, Qwen3-8B teacher, 3,000 unique training examples, and nine frozen checkpoints. Its full functional evaluation sets contain 500 MATH-500, 279 MATH-Perturb, and 600 Putnam items; the prespecified diagnostic also isolates 416 base MATH-500 retention items and 42 and 124 teacher-only items in the latter two sets. No two consecutive eligible intervals met its fitting/function discordance gate, leaving the observed acquisition-retention pattern as a diagnostic rather than a successful decoupling test. Fitting is $F_t = -\operatorname{mean} \mathrm{KL}(\text{student}\|\text{teacher})$ on held-out diagnostic prompts, aggregated by problem. The archived gate requires supported fitting improvement together with flat or negative MATH-Perturb recovery in at least two consecutive eligible checkpoint intervals. It uses 10,000 paired bootstrap replicates for the interval decisions, a three-percentage-point functional-decline threshold, and excludes ceiling or collapse-flagged intervals. The stronger gate additionally asks for a five-point decline, MATH-500 retention loss, and a same-direction Putnam result. The archived transition analysis records no gate-eligible discordant interval; a rise in the fitting point estimate alone therefore does not satisfy the gate. Table 43 provides the full nine-checkpoint accuracy, output-cap frequency, and mean response length for the three functional evaluation slices. The step-160 cap-status stratification in Table 44 shows that capped and uncapped items have different paired changes, but cap status is a later response property and does not identify the cause of a change. The completed FitFunc UGMath $K = 32$ endpoint check follows 96 fixed items with 3,072 responses at each endpoint: 1,820 are scored correct at step 0 and 1,816 at step 160, or 59.24% and 59.11% under the fixed denominator. Its paired change is -0.13 percentage points, with 13 items increasing, 15 decreasing, and 68 tied. This two-endpoint check is distinct from the structural-distance study below and is not an independent training replication of the Qwen3-4B result.

MATH-500 ($n = 500$)

Step	Acc.	Cap	Tokens
0	83.20	18.80	4401
5	81.20	20.80	4461
10	80.40	19.60	4602
20	71.20	35.80	6075
40	77.00	24.20	4890
60	78.00	23.40	4859
80	74.60	27.00	5027
120	76.40	26.20	4954
160	76.00	24.80	4961

MATH-Perturb ($n = 279$)

Step	Acc.	Cap	Tokens
0	63.08	18.64	9613
5	59.14	22.94	9741
10	59.86	26.88	10266
20	50.18	43.01	12208
40	55.91	30.11	10816
60	55.56	35.48	11161
80	56.99	33.69	11305
120	54.84	35.48	11479
160	54.48	36.56	11196

Putnam ($n = 600$)

Step	Acc.	Cap	Tokens
0	28.17	48.67	13688
5	26.67	52.33	13647
10	24.17	56.00	14005
20	13.83	80.00	15478
40	21.50	67.83	14688
60	18.00	72.33	15047
80	15.83	75.17	15038
120	17.00	72.67	14988
160	18.00	71.67	14837

Table 43: Accuracy, output-cap frequency, and mean response length across nine saved FitFunc checkpoints in three functional slices. These paths do not meet the prespecified persistent-discordance gate.

Slice	Cap at step 160	Items	Step 0 (%)	Step 160 (%)	Change (pp)	Token change
MATH-500	No	376	97.34	97.07	-0.27	+479
MATH-500	Yes	124	40.32	12.10	-28.23	+808
MATH-Perturb	No	177	80.79	85.88	+5.08	+495
MATH-Perturb	Yes	102	32.35	0.00	-32.35	+3472
Putnam	No	170	48.24	63.53	+15.29	-284
Putnam	Yes	430	20.23	0.00	-20.23	+1715

Table 44: FitFunc endpoint scores grouped by whether the step-160 response reached the output cap. Cap status is measured after training, so rows describe association.

The separate UGMath V5.2 structural Δp study completed a $K = 32$ stage with 6,144 generations. Only four families met its informative-family criterion, below the frozen minimum of eight, so the planned $K = 64$ continuation was not run. The stopped diagnostic does not establish a signed structural-distance effect.

I Post Hoc Calibration of the Main Direction Rule

The following analysis addresses the rule used for the main Llama results, separately from the archived Jeffreys and Fisher simulations in Appendix F. Its settings were recorded before this CPU computation, after the empirical results were known. It is a post hoc sensitivity analysis, not a preregistered validation study. No model responses are generated.

I.1 Match selection, budgets, prior, and decision threshold

For a fixed endpoint probability pair (p_0, p_1), independent D counts use eight attempts per endpoint. An item enters if exactly one endpoint has at least one success. Separate validation counts use 128 or 512 attempts and the Beta$(1, 1)$ prior. The original-direction probability is thresholded at 0.975 and 0.025, exactly as in Section 2.1. All selected Llama validation counts are binary, so the empirical missing-score branch is empty. Selection membership and validation counts are independent under this synthetic model; their finite count probabilities are multiplied and summed without fitting the grid to the observed candidates.

For equal endpoint counts n, the direction probability has the finite representation

$$P(p_1 > p_0 \mid c_0, c_1) = P(H \le c_1), \qquad H \sim \text{Hypergeom}(2n + 2, n + 1, c_0 + c_1 + 1), \tag{4}$$

where the hypergeometric parameters denote population size, number of successes, and sample size. This identity follows by ordering the two independent sets of $n + 1$ uniform variates whose relevant order statistics have the two Beta posteriors. We enumerate all count pairs. Ten boundary and interior fixtures agree with adaptive integration to within 8×10^{-16}; a moment-matched normal approximation is not used.

The baseline grid is $\{0, 0.001, 0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99, 1\}$. Requested shifts are 0, ± 0.005, ± 0.01, and ± 0.05, with probabilities clipped to $[0, 1]$. For the summary, each baseline receives 155 artificial items, giving 1,860 items in total. This matches only the empirical universe size. It does not recover its difficulty distribution or simulate a new trained model. Error under the null means any directional declaration; under a nonzero shift it means a declaration with the wrong sign. Each conditional rate weights the grid by its candidate-selection probability.

Requested shift (pp)	Validation n	Expected selected	Wrong direction | selected	Correct | selected
$+0.0$	128	187.84	3.876%	0.00%
$+0.0$	512	187.84	4.768%	0.00%
$+0.5$	128	205.18	1.262%	2.64%
$+0.5$	512	205.18	1.005%	5.91%
$+1.0$	128	221.93	0.827%	3.80%
$+1.0$	512	221.93	0.431%	16.15%
$+5.0$	128	336.53	0.031%	47.49%
$+5.0$	512	336.53	0.001%	87.29%

Table 45: Iid calibration of the main $K = 8$ selection and direction rule on 1,860 artificial items. Zero-shift rows measure false directions; positive-shift rows separate wrong signs from unresolved comparisons.

The null error rate among selected items is 3.88% for 128 validation responses and 4.77% for 512. Across the artificial universe, the corresponding expected false declarations are 7.28 and 8.96, and the probability of at least one exceeds 0.999 under independent items. The nominal itemwise rule consequently supplies no family error guarantee. With a positive half-point shift, correct-direction declaration rates among entrants are only 2.64% and 5.91%; for a five-point shift they rise to 47.49% and 87.29%. Those mixture-specific rates do not prescribe a universally sufficient validation budget. Every baseline and both shift directions are retained in the released grid.

I.2 Stress the response-independence assumption

Within each synthetic batch, responses share a latent probability drawn from a Beta distribution with mean p and intraclass correlation ρ. The resulting batch count is beta-binomial; independent

batch-count distributions are convolved to reach the required budget. We use $\rho \in \{0.01, 0.05\}$ and batch sizes 32 or 128, with the eight discovery attempts in one batch. Endpoint, batch, pool, and item latent variables are independent. The analyst nevertheless applies the original iid Beta posterior to the resulting counts.

Within-batch ρ	Batch size	Null error, $n = 128$	Null error, $n = 512$
0.00	32	3.88%	4.77%
0.01	32	6.98%	8.31%
0.05	32	19.00%	21.32%
0.01	128	16.24%	18.62%
0.05	128	40.31%	45.42%

Table 46: No-change calibration of the main posterior direction rule under synthetic within-batch correlation. Rates condition on $K = 8$ discovery selection; correlations are assumed, not estimated from the experiment.

Even $\rho = 0.01$ increases false directional declarations substantially when 128 responses share a batch. The correlation values and batch sizes are chosen stress settings, not estimates from generation logs. This check establishes a failure mode of the iid analysis under specified dependence; it does not establish that the recorded data have that dependence, quantify their effective sample size, or validate a correction for it. A larger response count improves nominal power while leaving model misspecification unresolved.

I.3 Multiplicity and magnitude in the observed cohort

Both stricter checks retain the same frozen list of 151 items. Fisher's exact one-sided tests compare the two endpoint success counts, with Holm correction over both directions of every item, giving 302 hypotheses per response pool. Under independent Bernoulli responses, response holdout permits this test conditional on the selected list. The second check raises the posterior direction threshold to $1 - 0.05/302$. It bounds posterior tail risk only under the specified prior and likelihood; it is not a second frequentist calibration. Neither adjustment repairs correlated responses.

Table 1 shows that nine of the 21 nominal W reversals survive both checks, while none of the nine nominal V reversals survive either. For the surviving W subset, the observed signed effects range from -17.77 to -6.05 points with median -9.18. Table 47 separates sign decisions from posterior evidence for effect magnitude. The thresholds of two, five, and ten points were chosen for sensitivity analysis after the original experiment, and are not established application-specific minima. In particular, five surviving reversals have pointwise posterior probability at least 0.975 of exceeding five points; those magnitude decisions have no extra Holm correction. Released item records contain endpoint counts, exact directional probabilities, adjusted test values, effect probabilities, and posterior quantiles.

Sign rule	Direction	Sign count	$\lvert\Delta\rvert > 2$ pp	$\lvert\Delta\rvert > 5$ pp	$\lvert\Delta\rvert > 10$ pp
Nominal Beta	Original	96	87	73	58
	Opposite	21	10	5	1
Fisher/Holm	Original	82	81	73	58
	Opposite	9	9	5	1

Table 47: W_{512} items whose Beta$(1, 1)$ posterior assigns at least 0.975 probability to a change larger than each stated threshold. Counts are pointwise magnitude checks within the indicated direction.

I.4 Recorded response partitions for the nine reversals

We reconstruct all 154,624 W membership rows, verify distinct request identities and complete sample indices, and recover 512 responses at every item and endpoint. The four partitions are defined by response hash rank, with 128 responses per endpoint in each. Their correct counts sum

exactly to the main W totals. They are not execution batches, worker assignments, or independent generation restarts.

Table 48 reports every partition for the nine Holm-selected reversals. All nine have a contrary observed sign in all four partitions, but only two have nominal opposite support in all four. Five have opposite support in the separate V measurement and four are uncertain there. We additionally omit each hash partition in turn, analyze the remaining 384 responses, and apply Holm over both signs of all 151 items anew. The nine-item retained counts are seven, seven, eight, and nine; six items survive all four omissions. This is post hoc sample-partition sensitivity, with overlapping reduced samples and item selection based on the full W data. It neither supplies four replications nor corrects unknown execution-batch dependence.

Item	Block 0	Block 1	Block 2	Block 3	Nominal	Omission	V
IFEval 2785	-7.03	-7.81	-5.47	-3.91	2/4	2/4	U
IFEval 2941	-20.31	-20.31	-15.62	-14.84	4/4	4/4	O
IFEval 3327	-7.81	-7.81	-6.25	-10.16	3/4	4/4	U
IFEval 3362	-9.38	-10.94	-6.25	-5.47	2/4	2/4	U
IFEval 3456	-4.69	-7.03	-5.47	-7.81	3/4	4/4	O
GSM8K 1094	-6.25	-8.59	-19.53	-7.03	1/4	3/4	O
GSM8K 1129	-9.38	-12.50	-16.41	-13.28	3/4	4/4	U
GSM8K 1145	-10.16	-10.94	-7.03	-8.59	4/4	4/4	O
GSM8K 1174	-19.53	-6.25	-11.72	-11.72	3/4	4/4	O

Table 48: Signed effects for nine Llama reversals in four disjoint 128-response hash partitions. Nominal and omission columns count opposing Beta and Holm decisions; V denotes the separate V_{128} decision.

I.5 Request-index segments and validation depth

The saved W sample indices permit a second partition that does not use response hashes. We form four consecutive index segments, 0–127, 128–255, 256–383, and 384–511, each with 128 responses at both endpoints. Sample index is a request identifier; it need not preserve wall-clock execution order or physical batches. All nine corrected reversals have an opposite observed sign in every segment. Omitting segments in order and repeating Holm across the entire 151-item family retains nine, seven, eight, and eight of the original nine; six survive all four omissions.

We also check whether correct-score proportions vary across segments within each of the 302 item-endpoint streams. Conditioning on its total correct count, the iid reference distribution allocates those successes over four groups of size 128 through a multivariate hypergeometric draw. The statistic is the Pearson sum for correct and noncorrect cells. We use 100,000 conditional draws with seed 2026092602, include ties in the upper tail, and compute Monte Carlo p values as $(1 + \#\{T^* \geq T\})/100001$. Degenerate all-correct or all-incorrect streams receive $p = 1$. Holm correction over the 302 streams rejects none. Failure to reject does not establish equivalence, response independence, or absence of within-segment dependence.

Table 49 reports the request-index depth analysis discussed in the main text. The numerical supplement provides per-item prefix counts, conditional-test values, omission results, and Holm decisions at depths 128, 256, 384, and 512. The 64-response row uses author-reported counts without a released response-membership record.

	Beta$(1,1)$, threshold 0.975			Fisher/Holm, 302 hypotheses		
W responses per endpoint	Original	Opposite	Unresolved	Original	Opposite	Unresolved
$64^{\dagger}$	34	3	114	13	0	138
128	78	9	64	56	0	95
256	86	13	52	75	5	71
384	92	16	43	81	8	62
512	96	21	34	82	9	60

Table 49: Direction decisions for 151 fixed Llama items as W depth grows. Rows 128 to 512 use saved request-index prefixes; $64^{\dagger}$ uses author-reported counts without a released response-membership record. Holm covers 302 one-sided hypotheses at each depth.

I.6 Missing outcomes and uncertainty in the Qwen comparison

The deterministic greedy screen uses a different decoding distribution from C/V. We therefore do not derive its admission probability from stochastic response probabilities. A component calculation instead conditions on an already fixed candidate with endpoint probabilities (p_0, p_1). It uses independent C_{128} and V_{128} counts, the same Beta prior and threshold, and the direction-specific least-favourable completion of unknown outcomes. Missing masks are independent of correctness, with endpoint missing rates $(0, 0)$, $(0.01, 0.01)$, $(0.05, 0.05)$, $(0.05, 0)$, or $(0, 0.05)$. Under this artificial assumption, a positive-direction check uses baseline success probability $p_0 + (1 - p_0)r_0$ and later success probability $p_1(1 - r_1)$ in its completed counts.

At $(p_0, p_1) = (0.1, 0.1)$, a positive declaration has probability 2.32% without missingness, 1.23% with one percent missing at each endpoint, and 0.067% with five percent missing. At $(0.1, 0.15)$ the corresponding rates are 22.06%, 14.99%, and 2.11%. For fixed truth and independent pools, the same rate applies to V conditional on C confirmation. The released table includes every grid point. This component calculation illustrates the conservatism of adverse completion; it does not model informative missingness, semantic scoring errors, or the full deterministic acquisition selector.

Sampling uncertainty is computed separately from unknown-score completion. For the frozen score-yield target, a response contributes one only when its frozen score is correct. With item membership held fixed, independent Beta$(c + 1, 129 - c)$ draws at each endpoint yield posterior distributions of the item-equal changes and the C95 minus R123 contrast. Table 5 displays the empirical changes, posterior means, and equal-tailed intervals from 100,000 draws with seed 20260926. C95 has interval $[-12.94, -11.17]$ percentage points, R123 has $[-0.69, +0.10]$, and their contrast has $[-12.73, -10.80]$. The posterior means are -12.06, -0.29, and -11.76 points. Here a posterior mean targets fixed-denominator correct yield; it is distinct from the conditional-on-resolved-score Beta mean in Table 4. This is response-model uncertainty for the fixed groups. Item resampling targets a different source of variation, while unknown-score bounds allow a different scoring completion. These three objects must not be substituted for one another, and none measures variation across training runs.

J Anonymous Review Materials and Evidence Scope

Table 50 indexes the anonymous numerical and response supplement. Its nine-item ledger includes public benchmark IDs, $D/V/W$ counts, effect quantiles, and adjusted tests. Request identities and hash partitions accompany all 9,216 endpoint scores for those items. These membership records contain no full answers.

Supplement file or directory	Evidence and scope
`data/NINE_REVERSE_ITEM_EVIDENCE.csv`	Nine-item counts, decisions, effect quantiles, and links to $D/V/W$.
`data/EMPIRICAL_MULTIPLICITY.csv`	Both signs of the full 151-item family, separately for V and W.
`data/W_ALL151_HASH_BLOCK_COUNTS.csv`	All four partitions for every item; reduced-sample Holm outputs are alongside this table.
`data/QWEN_FIXED_ITEM_RESPONSE_UNCERTAINTY.csv`	Empirical changes, posterior means, and interval endpoints for fixed groups.
`data/` and `inputs/`	Calibration, matching, selector replay, unified rules, recovered S responses, execution-worker checks, and frozen counts.
`metadata/KNOWN_PROVENANCE.json`	Available model, prompt, scorer, and protocol bindings, with explicit unavailable fields.
`human_review/`	The 52 available primary-V answers and a separate full 2,560-answer S packet; no completed human labels.
`code/` and `README.md`	CPU reproduction commands and input/output hashes, without local usernames or machine paths.

Table 50: Anonymous release files for reproducing the reported numerical analyses. The listed materials have the input coverage shown in each row.

The metadata distinguishes planning freezes from completed measurements and artifact identity from unverified training history. The reproduction tests all 151 items before extracting nine reversals and supplies request-index checks, score-change witnesses, and C95 concentration analyses. The review packets contain 52 primary-V answers and all 2,560 corresponding answers from the recovered S source, each with an empty judgment form. No human adjudication is claimed. The supplement also includes all 24,320 S endpoint texts and their frozen score links, permitting direct inspection of this separate measurement.